Capstone Project Report

Project 2

Document-based Question Answering System

(DocQA)

Submission by Team 42

Name	*OHR ID*
Varun Batra	*850072588*
Shaurya Singh	*850073054*
Ankitha Reddy Vancha	*850073073*
Sara Mehraj	*850073123*
Sai Dev Anirudh Thatode	*850073066*

Gen AI CoE Talent Academy

Developer Track

Batch 3

1. Problem Statement

In the context of information retrieval from diverse document formats such as PDFs and Standard Operating Procedures (SOPs) containing text and images, there exists a critical need to develop an automated AI-driven solution. This solution must accurately extract relevant information to answer user queries efficiently. Moreover, the system should be capable of handling and interpreting visual content embedded within these documents, thereby enhancing its ability to respond to queries based on both textual and graphical information. System also will be able to streamline data retrieval processes, enhance user interaction, and empower decision-making by providing access to thorough information.

2. Objectives

The project aims to develop an AI-driven solution to automate query responses by extracting precise information from documents, including text and images.

- **Develop AI-driven Information Extraction**: Execute models and algorithms that can reliably extract textual and visual data from SOPs and PDFs, providing efficient handling of a variety of content types and document formats.
- **Enable Query Answering Capability**: Develop and implement a system that leverages information extracted from documents (text and images) using a Retrieval-Augmented Generation (RAG) pipeline. This system aims to understand user queries effectively and provide precise responses based on retrieved document content.
- **Ensure Accuracy and Reliability**: Verify and improve the AI models to ensure high accuracy in information extraction and query response, minimizing mistakes and enhancing system reliability.
- **Provide Contextual Source Information**: Provide systems that will not only respond to inquiries but will also give comprehensive source data, like page numbers and document names, to improve tracking and transparency.
- **Enhance User Experience**: Enhance the user experience overall by allowing easy access to important data, cutting down on the time and effort needed for manual searches, and enabling easy interaction with the AI-powered system.

3. Approach and Logic

Our approach is to implement an AI-driven system for automated document querying and answering. It begins by processing a collection of PDF documents, extracting both textual and image-based content using OCR and computer vision techniques. The extracted information is indexed using dense embeddings generated by a transformer model. User queries are processed through a hybrid search mechanism combining dense and sparse search strategies. The most relevant document is selected and re-ranked based on query relevance using a conversational AI model. Finally, the system provides concise answers to queries along with the source document's name and page number. This approach aims to streamline information retrieval from diverse document types, enhancing decision-making and user experience by efficiently accessing critical data points.

4. Why this approach?

- **Document Diversity**: Oversees diverse document types including PDFs with text and images using PyPDF2 for text and tesseract LSTM for image OCR.
- **Document Processing and Parallelization**: Concurrent processing with Process Pool Executor optimizes document handling, improving efficiency and reducing processing time for large document sets.
- **Text Chunking and Embedding**: Text is split into chunks for efficient semantic embedding using Sentence Transformers, stored in ChromaDB for quick retrieval based on semantic similarity.
- **Hybrid Search**: Integrates dense (vector-based) and sparse (keyword-based) search methods for comprehensive document retrieval and optimization of RAG pipeline.
- **Re-ranking with Language Models**: Uses language models like GPT-4 to re-rank documents based on query relevance, ensuring the most pertinent information is presented first.
- **Efficiency and Scalability**: Employs multiprocessing with Process Pool Executor and concurrent.futures for efficient handling of large document volumes.
- **User-Centric Design**: Focuses on user interaction with an intuitive query interface and structured output (Excel), ensuring usability and practical utility.

5. User Guide

5.1 Installation Steps

1.Virtual Environment Setup: Create a virtual environment to manage dependencies.

a. Open a terminal or command prompt and run:

```
python -m venv env
```

b. Activate the virtual environment:

For Windows:

```
\env\Scripts\activate
```

For macOS and Linux:

```
source env/bin/activate
```

2. Install required libraries via the ‘requirements.txt’ file.

```
pip install -r requirements.txt
```

3. In the config file.

a. Update the path of tesseract:

```
which tesseract #command to retrieves the path on Linux system
Tesseract_CMD = path
```

b. Update OpenAI API key.

```
OPENAI_API_KEY = 'sk-' #Replace with your OpenAI API key.
```

5.2 Running the Application:

Python version used: 3.11.7.

1. Start the Application:

```
python main.py
```

2. Interact with the Application:

a. Enter your queries when prompted.

b. Type 'done' to finish entering queries.

3.View Results: The results will be displayed on the CLI along with saving it in output_file.xlsx in the data/documents directory.

6. Implementation

6.1 Python Libraries:

- **OpenCV-python (4.10.0.84):** Library for computer vision tasks such as image and video processing.
- **PyPDF2 (3.0.1):** Python library to handle PDF files, allowing manipulation like splitting, merging, and extracting text.
- **pdf2image (1.17.0)**: Converts PDF pages to images, facilitating image-based PDF processing.
- **pytesseract (0.3.10)**: Python wrapper for Google's Tesseract-OCR Engine, used for optical character recognition (OCR).
- **chromadb (0.5.5)**: A library for accessing and querying ChromaDB, a database of chromatin accessibility data.
- **scikit-learn (1.5.1)**: Machine learning library providing tools for data mining and analysis, built on NumPy, SciPy, and matplotlib.
- **pandas (2.2.2)**: Data analysis library offering data structures and operations for manipulating numerical tables and time series.
- **OpenAI (0.28)**: Python package for accessing OpenAI's API services, such as language models and other AI tools.
- **openpyxl (3.1.5)**: Python library to read/write Excel xlsx/xlsm/xltx/xltm files.
- **sentence-transformers (3.0.1)**: Library for generating and utilizing sentence embeddings using transformer models in this case ‘All-MiniLM-L6-v2'.
- **nltk (3.8.1)**: Natural Language Toolkit library for symbolic and statistical natural language processing (NLP) for English.

6.2 Methodology

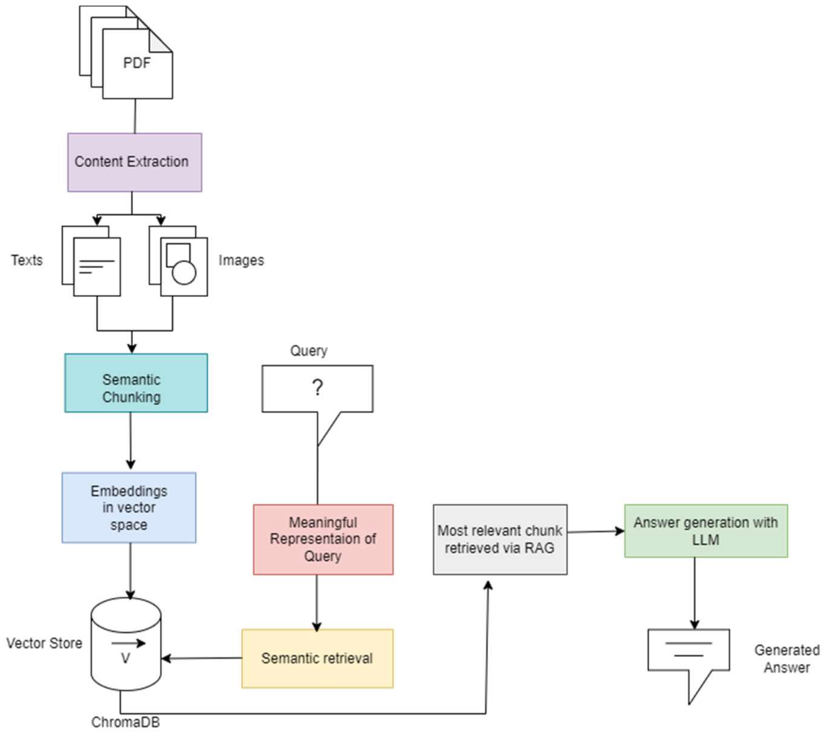

PDF
Content Extraction
Texts
Images
Semantic Chunking
Query
?
Embeddings in vector space
Meaningful Representaion of Query
Most relevant chunk retrieved via RAG
Answer generation with LLM
Vector Store
V
Semantic retrieval
Generated Answer
ChromaDB

Fig 1: Process Flow Diagram of Project

6.3 Steps

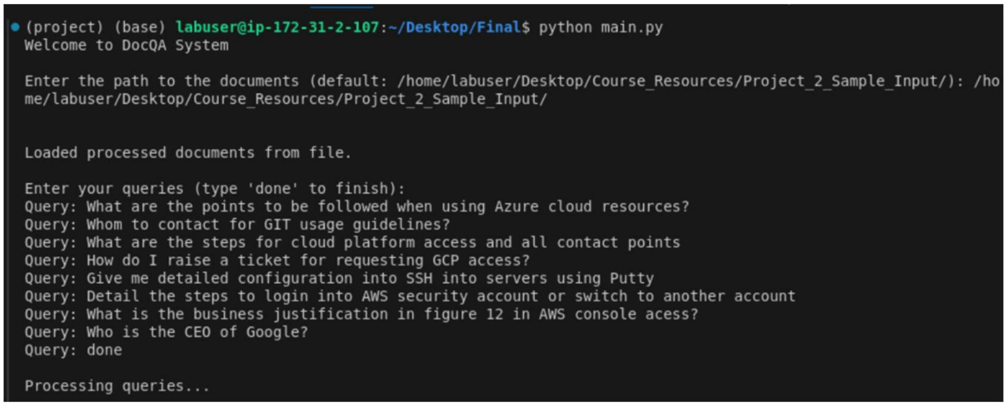
1. **Document Processing:**
   - PDF Text Extraction: Utilizes PdfReader from PyPDF2 (`extract_text_from_pdf`) to extract text from PDF documents.
   - Image OCR: Uses `pdf2image` (`convert_from_path`) and OpenCV (`preprocess_image`, `extract_text_from_images_in_pdf`) with Pytesseract (`pytesseract.image_to_string`) with custom_config = **r'--oem 1 --psm 6'** to initialize LSTM engine for OCR to extract text from images in PDFs.
   - Text Preprocessing and Chunking:
     - Text Preprocessing: Implemented in `preprocess_image` function using OpenCV for image preprocessing and Pytesseract for OCR.
     - Chunking: Done in `split_text_to_chunks` function to segment text into manageable chunks for indexing.
2. **Semantic Embedding and Vector Store Creation:**

- Sentence Embedding: Implemented using Sentence Transformers (`create_vector_store` function) to convert text chunks into semantic embeddings.
- Vector Store: Uses ChromaDB (`create_collection`, `add`) to store text chunks with their embeddings for efficient retrieval.

3. **Hybrid Search Mechanism:**
    - Dense and Sparse Search: Implementing RAG optimizations in `hybrid_search` function to perform both vector-based dense search and keyword-based sparse search for retrieving relevant documents.
4. **Document Re-ranking:**
    - Relevance Scoring: Uses OpenAI's GPT model (`re_rank_documents`) to re-rank documents based on relevance to user queries, considering textual content and user feedback.
5. **Answer Extraction and Presentation:**
    - Answer Generation: Implemented in `get_answer_from_documents` function using GPT-4 to generate answers based on the most relevant document retrieved from hybrid search.
    - Source Identification: The find_page_number function uses cosine similarity to identify the source document and specific page where the answer was found, ensuring accurate source attribution based on query relevance.
6. **Output**
    - Excel Output: Implemented in `save_to_excel` function to save query results including question, answer, source file, and page number into an Excel spreadsheet.

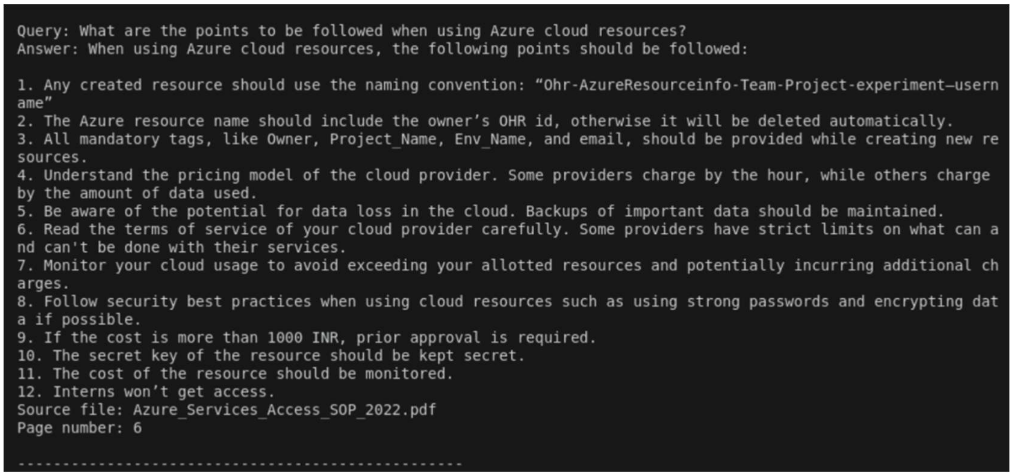
7. Result And Output:

A screenshot showing the system's accurate responses to valid user queries, along with corresponding source document names and page numbers.



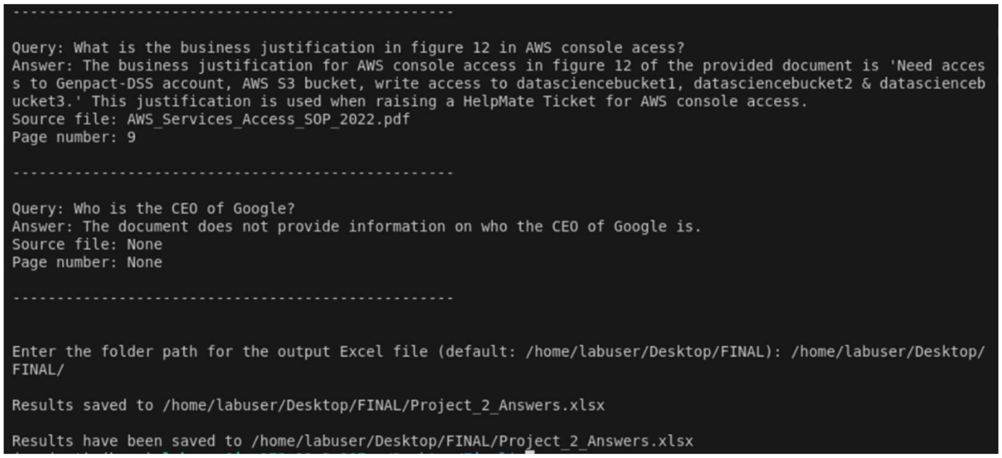
```
● (project) (base) labuser@ip-172-31-2-107:~/Desktop/Final$ python main.py
 Welcome to DocQA System

 Enter the path to the documents (default: /home/labuser/Desktop/Course_Resources/Project_2_Sample_Input/): /ho
 me/labuser/Desktop/Course_Resources/Project_2_Sample_Input/


 Loaded processed documents from file.

 Enter your queries (type 'done' to finish):
 Query: What are the points to be followed when using Azure cloud resources?
 Query: Whom to contact for GIT usage guidelines?
 Query: What are the steps for cloud platform access and all contact points
 Query: How do I raise a ticket for requesting GCP access?
 Query: Give me detailed configuration into SSH into servers using Putty
 Query: Detail the steps to login into AWS security account or switch to another account
 Query: What is the business justification in figure 12 in AWS console acess?
 Query: Who is the CEO of Google?
 Query: done

 Processing queries...
```

Fig 2. Input User Query and Data Files.



```
Query: What are the points to be followed when using Azure cloud resources?
Answer: When using Azure cloud resources, the following points should be followed:

1. Any created resource should use the naming convention: “Ohr-AzureResourceinfo-Team-Project-experiment—usern
ame”
2. The Azure resource name should include the owner’s OHR id, otherwise it will be deleted automatically.
3. All mandatory tags, like Owner, Project_Name, Env_Name, and email, should be provided while creating new re
sources.
4. Understand the pricing model of the cloud provider. Some providers charge by the hour, while others charge
by the amount of data used.
5. Be aware of the potential for data loss in the cloud. Backups of important data should be maintained.
6. Read the terms of service of your cloud provider carefully. Some providers have strict limits on what can a
nd can't be done with their services.
7. Monitor your cloud usage to avoid exceeding your allotted resources and potentially incurring additional ch
arges.
8. Follow security best practices when using cloud resources such as using strong passwords and encrypting dat
a if possible.
9. If the cost is more than 1000 INR, prior approval is required.
10. The secret key of the resource should be kept secret.
11. The cost of the resource should be monitored.
12. Interns won’t get access.
Source file: Azure_Services_Access_SOP_2022.pdf
Page number: 6

--------------------------------------------------
```

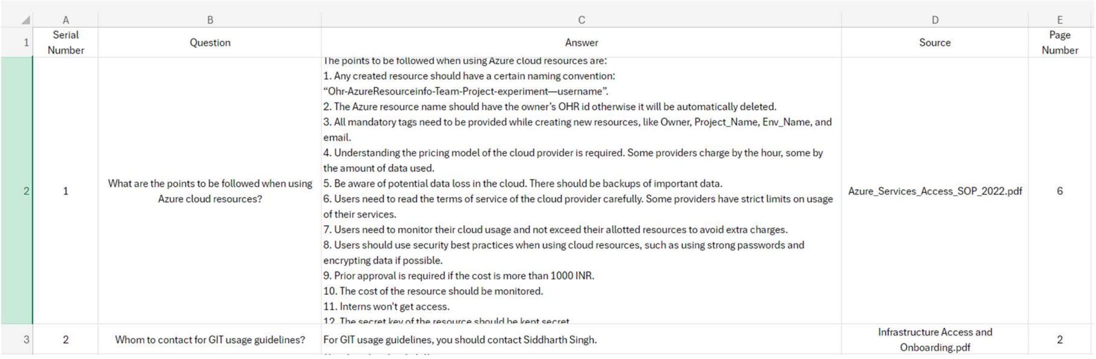
Fig 3. System's response to first user query.



```
--------------------------------------------------

Query: What is the business justification in figure 12 in AWS console acess?
Answer: The business justification for AWS console access in figure 12 of the provided document is 'Need acces
s to Genpact-DSS account, AWS S3 bucket, write access to datasciencebucket1, datasciencebucket2 & datascienceb
ucket3.' This justification is used when raising a HelpMate Ticket for AWS console access.
Source file: AWS_Services_Access_SOP_2022.pdf
Page number: 9

--------------------------------------------------

Query: Who is the CEO of Google?
Answer: The document does not provide information on who the CEO of Google is.
Source file: None
Page number: None

--------------------------------------------------


Enter the folder path for the output Excel file (default: /home/labuser/Desktop/FINAL): /home/labuser/Desktop/
FINAL/

Results saved to /home/labuser/Desktop/FINAL/Project_2_Answers.xlsx

Results have been saved to /home/labuser/Desktop/FINAL/Project_2_Answers.xlsx
```

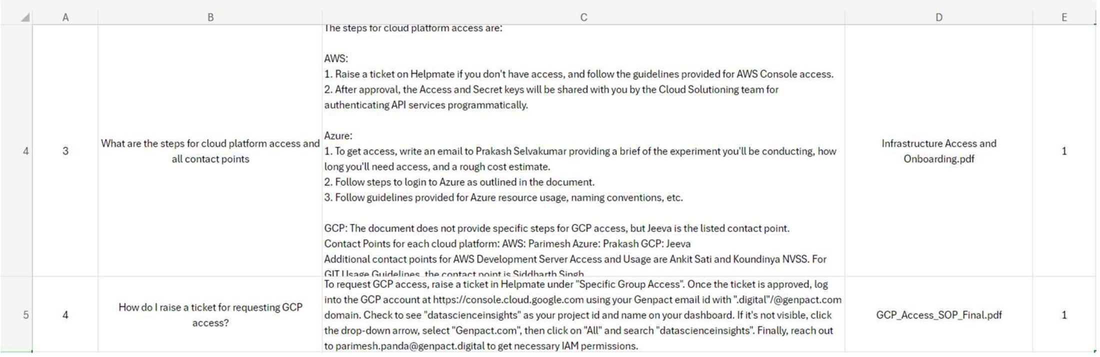
Fig 4. System's response to image related query and invalid query

Below screenshot showing the system's accurate responses to valid user queries, along with corresponding source document names and page numbers in the output excel file.



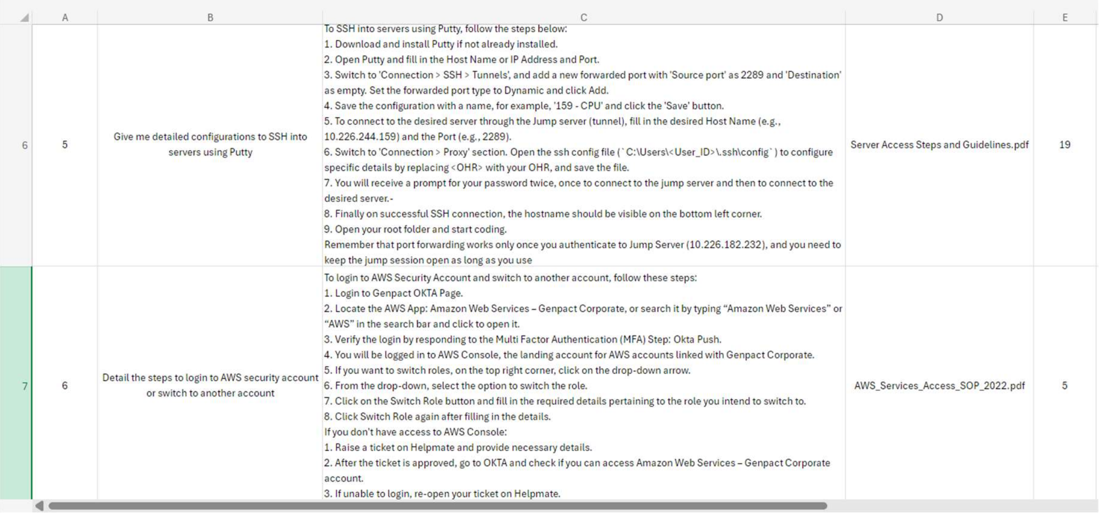
	A	B	C	D	E
1	Serial Number	Question	Answer	Source	Page Number
2	1	What are the points to be followed when using Azure cloud resources?	The points to be followed when using Azure cloud resources are: 1. Any created resource should have a certain naming convention: "Ohr-AzureResourceinfo-Team-Project-experiment—username". 2. The Azure resource name should have the owner's OHR id otherwise it will be automatically deleted. 3. All mandatory tags need to be provided while creating new resources, like Owner, Project_Name, Env_Name, and email. 4. Understanding the pricing model of the cloud provider is required. Some providers charge by the hour, some by the amount of data used. 5. Be aware of potential data loss in the cloud. There should be backups of important data. 6. Users need to read the terms of service of the cloud provider carefully. Some providers have strict limits on usage of their services. 7. Users need to monitor their cloud usage and not exceed their allotted resources to avoid extra charges. 8. Users should use security best practices when using cloud resources, such as using strong passwords and encrypting data if possible. 9. Prior approval is required if the cost is more than 1000 INR. 10. The cost of the resource should be monitored. 11. Interns won't get access. 12. The secret key of the resource should be kept secret	Azure_Services_Access_SOP_2022.pdf	6
3	2	Whom to contact for GIT usage guidelines?	For GIT usage guidelines, you should contact Siddharth Singh.	Infrastructure Access and Onboarding.pdf	2

Fig 5. System's response for first two user query in Excel.



	A	B	C	D	E
4	3	What are the steps for cloud platform access and all contact points	The steps for cloud platform access are: AWS: 1. Raise a ticket on Helpmate if you don't have access, and follow the guidelines provided for AWS Console access. 2. After approval, the Access and Secret keys will be shared with you by the Cloud Solutioning team for authenticating API services programmatically. Azure: 1. To get access, write an email to Prakash Selvakumar providing a brief of the experiment you'll be conducting, how long you'll need access, and a rough cost estimate. 2. Follow steps to login to Azure as outlined in the document. 3. Follow guidelines provided for Azure resource usage, naming conventions, etc. GCP: The document does not provide specific steps for GCP access, but Jeeva is the listed contact point. Contact Points for each cloud platform: AWS: Parimesh Azure: Prakash GCP: Jeeva Additional contact points for AWS Development Server Access and Usage are Ankit Sati and Koundinya NVSS. For GIT Usage Guidelines, the contact point is Siddharth Singh	Infrastructure Access and Onboarding.pdf	1
5	4	How do I raise a ticket for requesting GCP access?	To request GCP access, raise a ticket in Helpmate under "Specific Group Access". Once the ticket is approved, log into the GCP account at https://console.cloud.google.com using your Genpact email id with ".digital"/@genpact.com domain. Check to see "datascienceinsights" as your project id and name on your dashboard. If it's not visible, click the drop-down arrow, select "Genpact.com", then click on "All" and search "datascienceinsights". Finally, reach out to parimesh.panda@genpact.digital to get necessary IAM permissions.	GCP_Access_SOP_Final.pdf	1

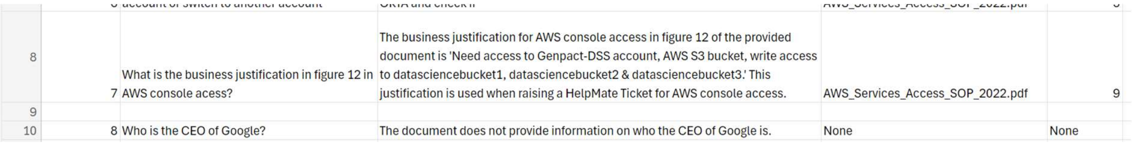
Fig 6. System's response for 3 & 4 user query in Excel.



	A	B	C	D	E
6	5	Give me detailed configurations to SSH into servers using Putty	To SSH into servers using Putty, follow the steps below: 1. Download and install Putty if not already installed. 2. Open Putty and fill in the Host Name or IP Address and Port. 3. Switch to 'Connection > SSH > Tunnels', and add a new forwarded port with 'Source port' as 2289 and 'Destination' as empty. Set the forwarded port type to Dynamic and click Add. 4. Save the configuration with a name, for example, '159 - CPU' and click the 'Save' button. 5. To connect to the desired server through the Jump server (tunnel), fill in the desired Host Name (e.g., 10.226.244.159) and the Port (e.g., 2289). 6. Switch to 'Connection > Proxy' section. Open the ssh config file (` C:\Users\<User_ID>\.ssh\config `) to configure specific details by replacing <OHR> with your OHR, and save the file. 7. You will receive a prompt for your password twice, once to connect to the jump server and then to connect to the desired server.- 8. Finally on successful SSH connection, the hostname should be visible on the bottom left corner. 9. Open your root folder and start coding. Remember that port forwarding works only once you authenticate to Jump Server (10.226.182.232), and you need to keep the jump session open as long as you use	Server Access Steps and Guidelines.pdf	19
7	6	Detail the steps to login to AWS security account or switch to another account	To login to AWS Security Account and switch to another account, follow these steps: 1. Login to Genpact OKTA Page. 2. Locate the AWS App: Amazon Web Services – Genpact Corporate, or search it by typing "Amazon Web Services" or "AWS" in the search bar and click to open it. 3. Verify the login by responding to the Multi Factor Authentication (MFA) Step: Okta Push. 4. You will be logged in to AWS Console, the landing account for AWS accounts linked with Genpact Corporate. 5. If you want to switch roles, on the top right corner, click on the drop-down arrow. 6. From the drop-down, select the option to switch the role. 7. Click on the Switch Role button and fill in the required details pertaining to the role you intend to switch to. 8. Click Switch Role again after filling in the details. If you don't have access to AWS Console: 1. Raise a ticket on Helpmate and provide necessary details. 2. After the ticket is approved, go to OKTA and check if you can access Amazon Web Services – Genpact Corporate account. 3. If unable to login, re-open your ticket on Helpmate.	AWS_Services_Access_SOP_2022.pdf	5

Fig 7. System's response for 5& 6 user query in Excel.

This section showcases the system's handling of various queries, including valid and invalid inputs. It includes screenshots of responses to nonsensical queries and demonstrates the system's ability to accurately extract and respond to data from images, ensuring comprehensive query resolution.



	6	account or switch to another account	OKTA and check if	AWS_Services_Access_SOP_2022.pdf	5
8	7	What is the business justification in figure 12 in AWS console acess?	The business justification for AWS console access in figure 12 of the provided document is 'Need access to Genpact-DSS account, AWS S3 bucket, write access to datasciencebucket1, datasciencebucket2 & datasciencebucket3.' This justification is used when raising a HelpMate Ticket for AWS console access.	AWS_Services_Access_SOP_2022.pdf	9
9					
10	8	Who is the CEO of Google?	The document does not provide information on who the CEO of Google is.	None	None

Fig 8. System's response for 7 & 8 user query in Excel.

8. Key Findings

- **Efficient Reprocessing**: Once documents are processed, the system avoids redundant reprocessing by utilizing pre-processed data. This means that subsequent queries can be answered based on the initial processing, thereby optimizing resource use and processing time.
- **Enhanced Efficiency**: Rapid extraction and processing of text and images from documents by implementing parallelization thereby reducing manual search time.
- **Improved Accuracy**: AI models ensure precise and relevant query responses.
- **Streamlined Query Handling**: Hybrid search methods effectively retrieve and rank documents.

9. Future Scope

Enhance the AI-driven document retrieval system by implementing robust batch processing for efficient handling of large volumes, integrating advanced OCR and NLP models for seamless multilanguage support, developing a user-friendly GUI for intuitive document management and query interaction, and enabling real-time streaming capabilities for dynamic data ingestion and prompt query responses. These enhancements aim to elevate operational efficiency, user experience, and system adaptability across diverse information retrieval scenarios.

10. Conclusion

In conclusion, the project addresses the need for automating document-based question answering with AI technologies. It efficiently extracts information from diverse formats like PDFs and SOPs, improving accuracy and retrieval efficiency. The system's use of OCR, NLP, and dense embeddings ensures reliable responses, while features like batch processing, multilanguage support, and real-time streaming enhance usability and scalability. This comprehensive approach streamlines data access and decision-making, laying a solid foundation for future advancements in AI-driven document management and user interaction.