ChatPDF

CHAT WITH PDF
Using Hugging Face

My Project Content

https://drive.google.com/drive/folders

10 Slide

Created By: Hossam Fares

ChatPDF

PRESENTATION

CONTENT LIST

01 Problem Definition

02 Solutions

03 Models From HF 🤗

04 Pipeline

05 User Interface

ChatPDF

01 Problem Definition

01 Users often need to extract specific information from large volumes of PDF documents.

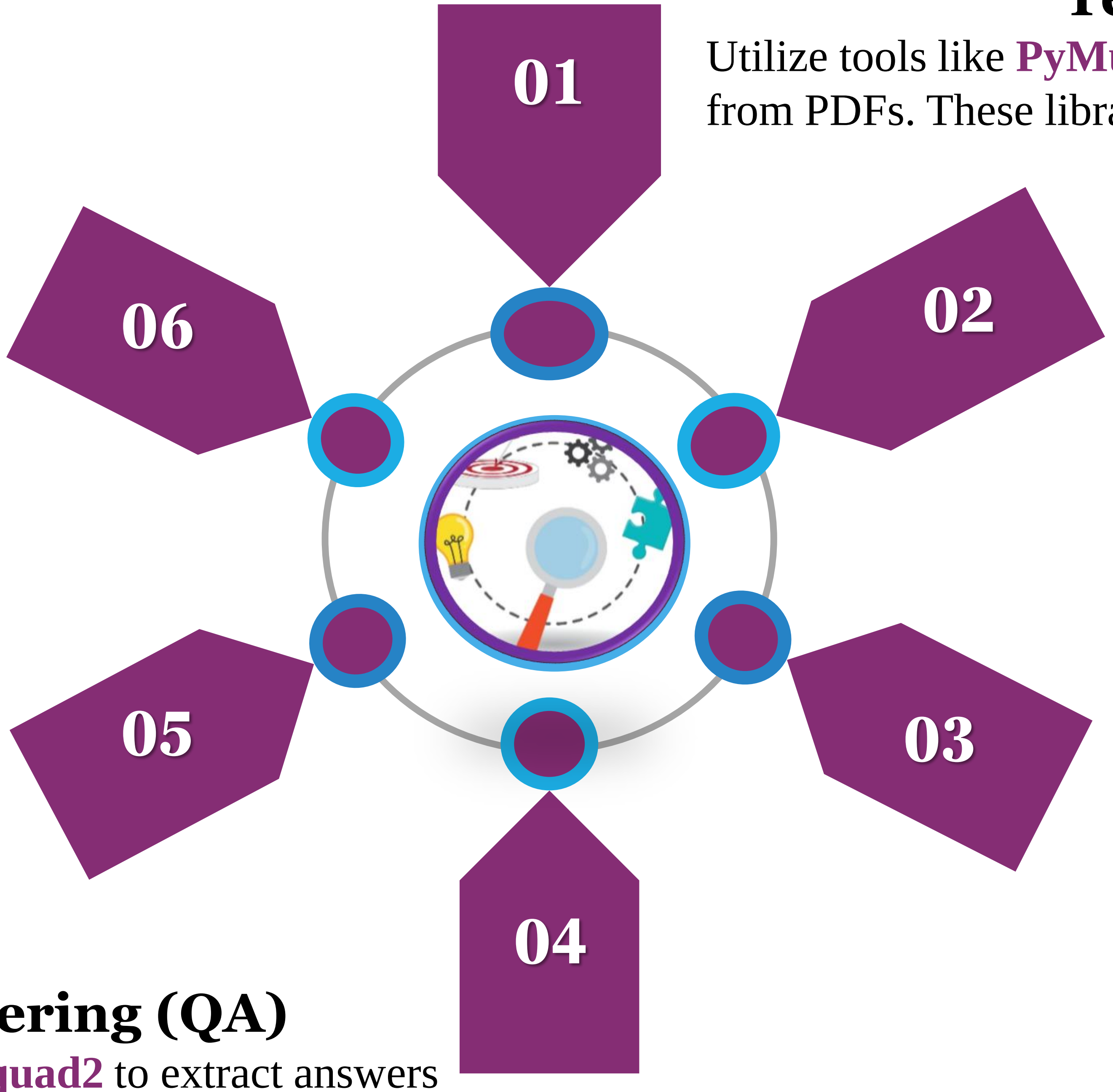
02 Traditional QA systems struggle with PDF formats, which may contain unstructured (text) data.

03 User queries may contain grammatical or spelling errors, leading to inaccurate or incomplete results from the QA system

04 Simply providing short answers may not be sufficient. Users might require detailed, well-structured explanations or descriptions derived from the extracted information

02 Solutions

ChatPDF



Text Extraction from PDFs

Utilize tools like **PyMuPDF**, **pdfminer**, or **PDFPlumber** to extract text data from PDFs. These libraries can handle both structured and unstructured content.

Text Splitting for Context Management

Use **LangChain's RecursiveCharacterTextSplitter** to divide extracted text into manageable chunks for efficient processing while preserving context.

Vector Search with FAISS

1- Use **FAISS from langchain.vectorstores.faiss** to index the text chunks.
2- Convert text chunks into embeddings using **HuggingFaceEmbeddings (sentence-transformers/all-MiniLM-L6-v2)** for similarity search.
3- Retrieve the most relevant passages based on the translated English query.

Question Answering (QA)

- Use **deepset/roberta-base-squad2** to extract answers from the relevant passages.
- **The QA model works in English, ensuring accurate and contextually relevant answers.**

Answer Translation

Translate the QA result from English back into the original language of the query using MarianMTModel and MarianTokenizer.
Provide the answer in both English and the user's original language.

Detailed Answer Generation

Use **GPT-2** for generating detailed explanations in English based on the QA output.
Translate the detailed answer back into the original language for users needing extended information.

02 Solution Cont.

ChatPDF

✓ Multilingual Query Translation

Leverage MarianMTModel and MarianTokenizer from transformers to handle multilingual queries.

Translate the user query from the original language to English

✓ Grammar and spelling correction

Leverage spelling-correction-english-base and grammar_error_correcter_v1 from transformers to handle error in queries.

✓ Summarization

Apply facebook/bart-large-cnn to summarize the detailed answers in English, then translate the summary back into the original language for concise responses.

03

Models From HF 🤗

ChatPDF

MarianMTModel

Multilingual translation for handling queries in various languages.

all-MiniLM-L6-v2

Embedding model for semantic search and similarity comparison.

RoBERTa (SQuAD2)

Question-answering model for extracting relevant answers from text.

spelling-correction and grammar_error_correcter

For correct query

GPT2

for generating detailed explanations in English based on the QA output.

BART

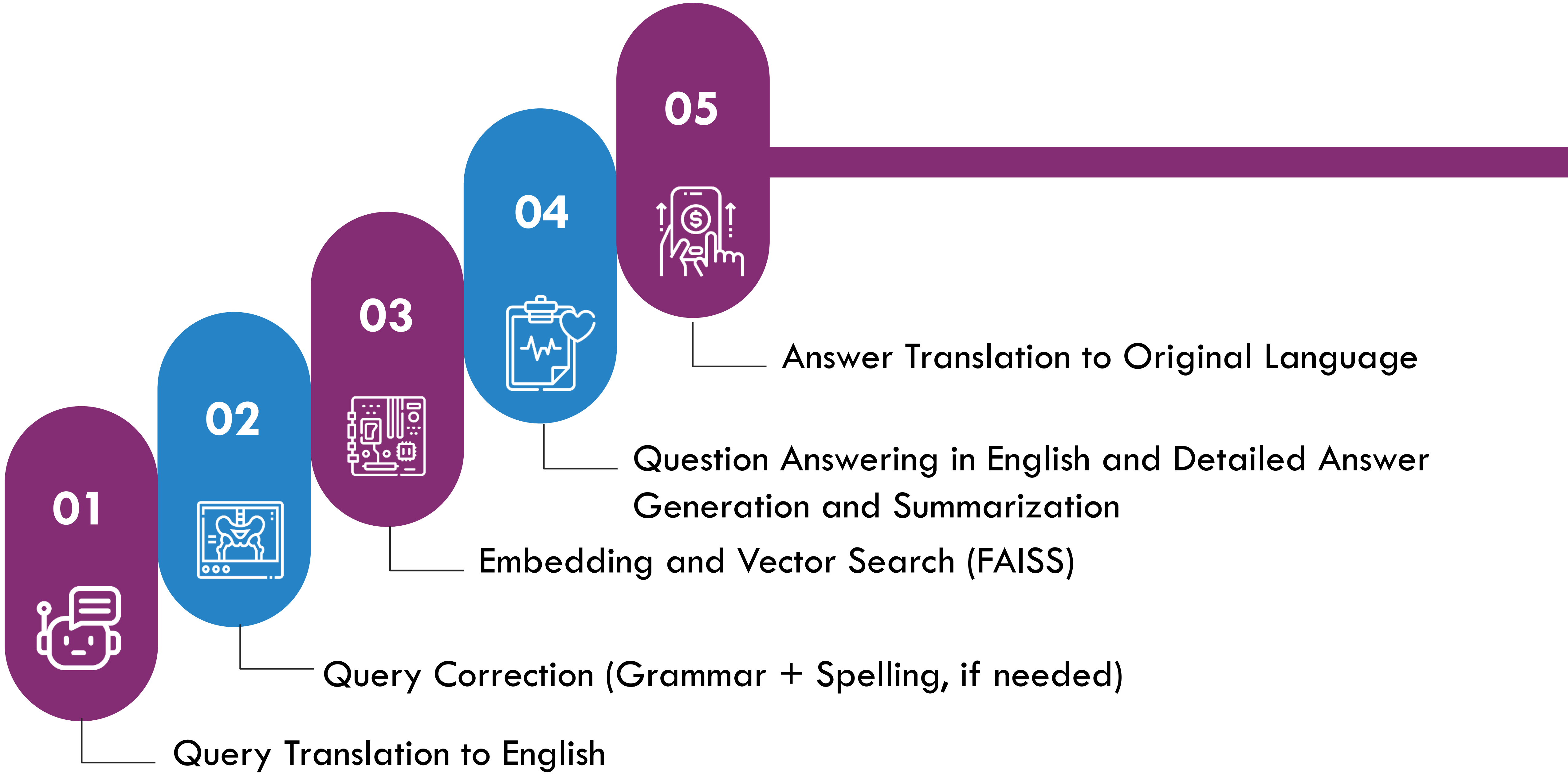
Summarization model for condensing detailed answers into concise formats.

Hugging Face

04 Pipeline

ChatPDF

Workflow Automation and Integration



01 Query Translation to English

02 Query Correction (Grammar + Spelling, if needed)

03 Embedding and Vector Search (FAISS)

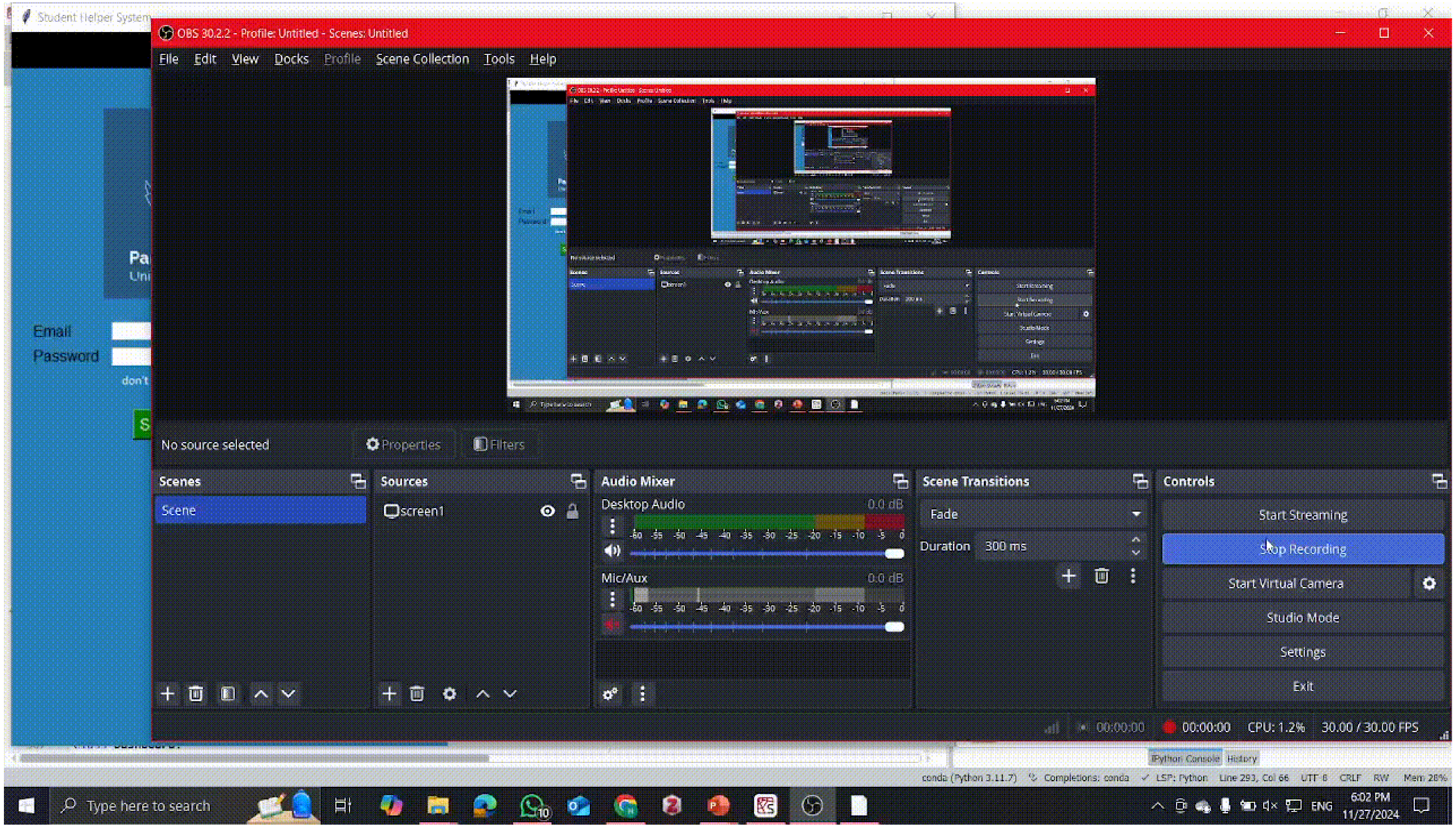
04 Question Answering in English and Detailed Answer Generation and Summarization

05 Answer Translation to Original Language

05

User Interface

ChatPDF



ChatPDF

Any Question??

gettyimages
Credit: Paul Gilligan
77379304

ChatPDF

Thank You...!

End