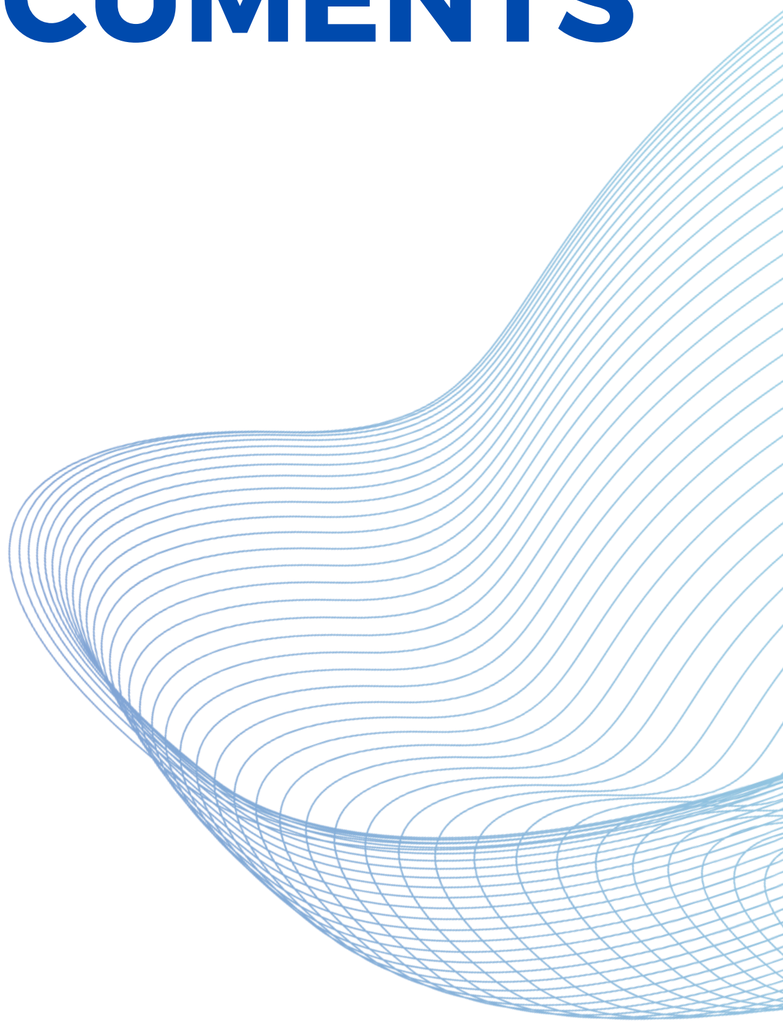
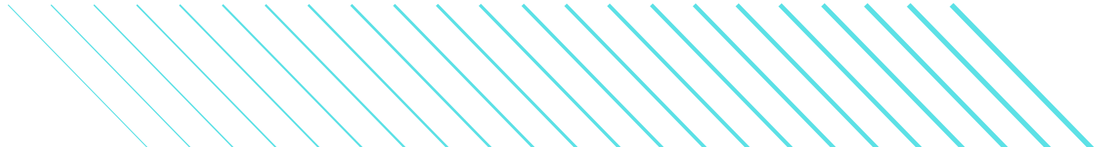
DS-AIBDA

CHAT WITH SCIENTIFIC DOCUMENTS USING LLMS

Gaurav Boob
42209





PROBLEM STATEMENT

Research and scientific endeavours heavily rely on the analysis and dissemination of information contained within various document formats.

While PDF remains a prevalent format for sharing research findings, scientists, academics, and professionals often utilize a diverse array of formats such as DOC, DOCX, TEX, and PPT for creating and presenting their work.

However, existing solutions primarily cater to parsing PDF documents, leaving a gap in efficiently handling these alternative formats.

INTRODUCTION

Documents are fundamental to research and professional communication, but existing solutions primarily focus on parsing PDF files. This leaves a gap in handling diverse formats like DOCX, TEX, and PPT, hindering efficient information extraction.

Our project aims to develop a robust framework to parse multiple formats beyond PDF, integrated with language models for intuitive user interaction and query answering.

This will streamline document analysis and knowledge dissemination, fostering a more efficient ecosystem for research and collaboration.

USE-CASES OF LLMS

DATA ANALYSIS

- LLMS can be used to analyze large datasets in scientific research.
- It can help identify patterns, correlations, and trends in the data.

EXPERIMENTAL DESIGN

- LLMS can assist in designing experiments for scientific research.
- It can help determine the appropriate sample size, control variables, and experimental conditions.

LITERATURE REVIEW

- LLMS can streamline the process of conducting a literature review.
- It can help researchers find relevant articles, extract key information, and organize references.

COLLABORATION

- LLMS can facilitate collaboration among researchers in scientific research.
- It can provide a centralized platform for sharing data, documents, and findings.

APPROACH

Document Parsing and Preprocessing:

- For text files (.txt), straightforward reading can be done using Python's built-in file handling.
- For PowerPoint (.pptx) and Pdf (.pdf) files, we have used Convert-api.
- Utilised libraries such as docx for parsing data from a docx file.
- for latex (.tex) files we are sending the code to the model.
- Preprocess the extracted text data to remove any irrelevant information, such as metadata , formatting artefacts or stop words.

Language Model Integration:

- Integrated Gemini 1.0 Pro, my chosen language model, for natural language understanding tasks and processing user queries.

APPROACH

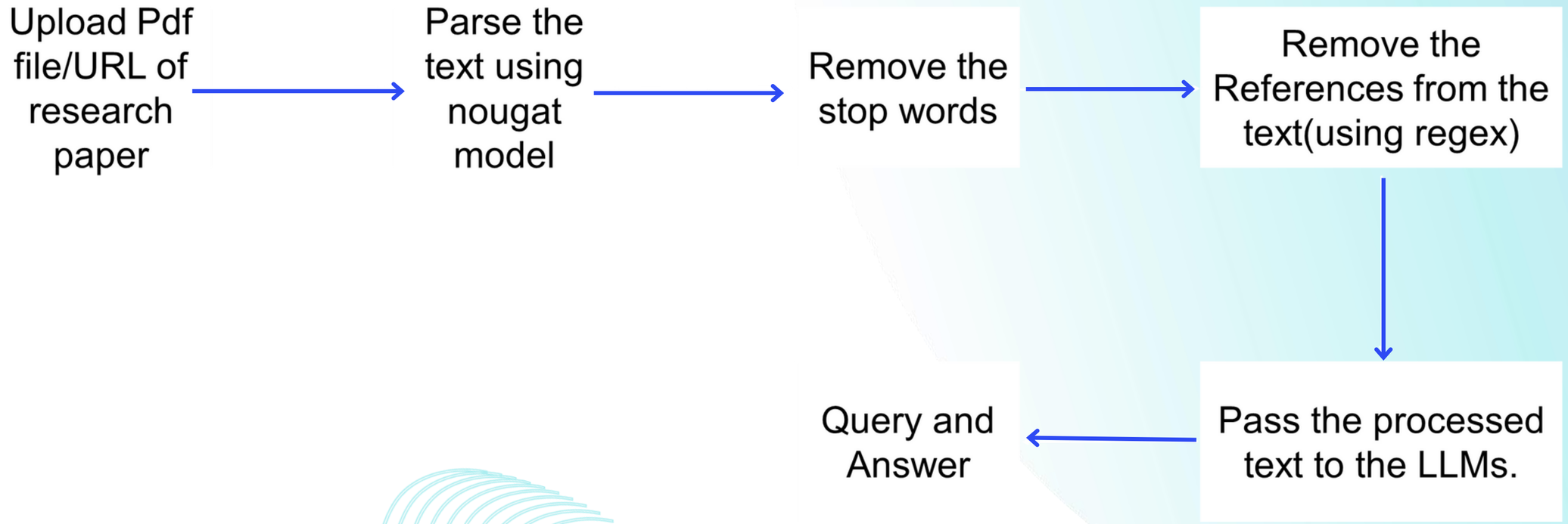
User Interface Development:

- Developed a web-based interface using Streamlit and Django frameworks to provide a user-friendly experience.
- Used tiktoken for authentication and authorization functionalities if required.
- Implemented interactive features allowing users to upload documents and input queries.

Document Analysis and Query Answering:

- Used Gemini 1.0 Pro for document analysis, extracting key insights, generating summaries and answering users query.

NOUGAT PIPELINE FLOW



ADVANTAGES

01

Efficient Research

LLMS allows researchers to quickly find relevant information within scientific documents through chat-based queries.

02

Natural Language Understanding

With LLMS, users can converse with scientific documents using natural language, making it easier to understand complex concepts.

03

Collaboration and Knowledge Sharing

LLMS enables seamless collaboration and knowledge sharing among researchers, allowing them to discuss and exchange ideas directly within the documents.

04

Time-saving

By chatting with scientific documents, researchers can save time by quickly extracting key information instead of manually searching through lengthy documents.

WHY SHOULD YOU CHOOSE THIS SOLUTION?

1. **Comprehensive Format Support**

This solution offers support for a wide range of document formats, including PDF, TXT, PPTX,Tex and DOCX. This ensures versatility and flexibility, allowing users to work with documents in their preferred formats without constraints

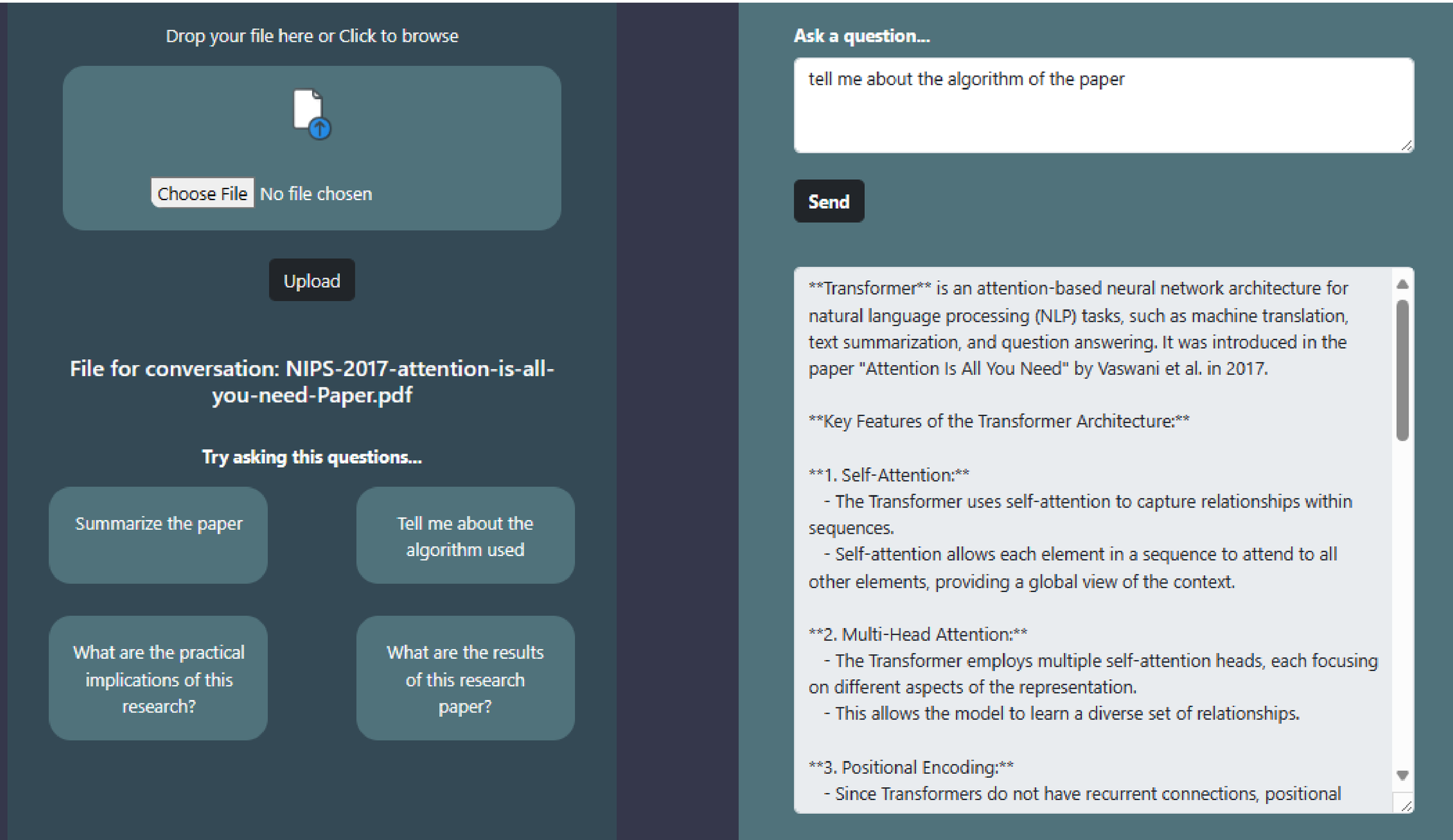
2. **Advanced Language Model Integration**

Integratation of Gemini 1.0 Pro, a powerful language model, into the solution enables advanced natural language understanding capabilities, including document analysis, summarization, and query answering. Users can leverage Gemini 1.0 Pro to gain valuable insights from their documents quickly and accurately.

3. **User-Friendly Interface**

The solution features a user-friendly web-based interface developed using Streamlit and Django frameworks. This interface makes it easy for users to upload documents, input queries, and interact with the system seamlessly. The intuitive design enhances user experience and productivity.

UI DESIGN


Drop your file here or Click to browse
Choose File
No file chosen
Upload
File for conversation: NIPS-2017-attention-is-all-you-need-Paper.pdf
Try asking this questions...
Summarize the paper
Tell me about the algorithm used
What are the practical implications of this research?
What are the results of this research paper?
Ask a question...
tell me about the algorithm of the paper
Send
Transformer is an attention-based neural network architecture for natural language processing (NLP) tasks, such as machine translation, text summarization, and question answering. It was introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017.
Key Features of the Transformer Architecture:
1. Self-Attention:
- The Transformer uses self-attention to capture relationships within sequences.
- Self-attention allows each element in a sequence to attend to all other elements, providing a global view of the context.
2. Multi-Head Attention:
- The Transformer employs multiple self-attention heads, each focusing on different aspects of the representation.
- This allows the model to learn a diverse set of relationships.
3. Positional Encoding:
- Since Transformers do not have recurrent connections, positional

THANK YOU