CHENNAI INSTITUTE OF TECHNOLOGY
(Autonomous)

PROJECT TITLE:

PDF CHATBOT WITH GENAI

A CORE COURSE PROJECT REPORT

Submitted By

THRISHA VIJAYAKUMAR

22AM059

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

DEPARTMENT OF CSE
(Artificial Intelligence & Machine Learning)
CHENNAI INSTITUTE OF TECHNOLOGY
(Autonomous)
Sarathy Nagar, Kundrathur, Chennai-600069

OCT / NOV – 2024

Vision of the Institute:

To be an eminent centre for Academia, Industry and Research by imparting knowledge, relevant practices and inculcating human values to address global challenges through novelty and sustainability.

Mission of the Institute:

IM1.To creates next generation leaders by effective teaching learning methodologiesand instill scientific spark in them to meet the global challenges.

IM2.To transform lives through deployment of emerging technology, novelty and sustainability.

IM3.To inculcate human values and ethical principles to cater the societal needs.

IM4.To contributes towards the research ecosystem by providing a suitable, effectiveplatform for interaction between industry, academia and R & D establishments.

CHENNAI INSTITUTE OF TECHNOLOGY
(Autonomous)

DEPARTMENT OF CSE
(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Vision of the Department:

The vision of the Department of Artificial Intelligence and Machine Learning is to impart quality education and produce high quality, creative and ethical engineers, in still professionalism, enhance students' problem solving skills in the domain of artificial intelligence and Machine Learning to emerge as a premier center for education and research in Artificial Intelligence and Machine Learning in transforming student into innovative professionals of contemporary and future technologies to cater the global needs of human resources for IT industries.

Mission of the Department:

DM1: To provide skill -based education to master the st udents in problem solving and analytical skills to enhance their niche expertise in the field Artificial Intelligence and Machine Learning .

DM2: To explore opportunities for skill development in the application of Artificial Intelligence and Machin e learning among rural and under privileged population .

DM3: Transform professionals into technically competent thr ough research -based projects in the emerging areas of Artificial Intelligence and Machine Learning and socially resp onsib le.

DM4: To impart quality and value -based education and contribute tow ards the innovation of computing system, data science to raise satisfaction level of all stakeholders .

CERTIFICATE

This is to certify that the "**Core Course Project**" Submitted by **THRISHA VIJAYAKUMAR (22AM059)** is a work done by him/her and submitted during **2024-2025** academic year, in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF ENGINEERING** in **DEPARTMENT OF CSE(Artificial Intelligence & Machine Learning),** at Chennai Institute of Technology.

Dr.P.KARTHIKEYAN
Project Coordinator
(Name and Designation)

Internal Examiner

DR.R.GOWRI
Head of the Department
(Name and Designation)

External Examiner

ACKNOWLEDGEMENT

We express our gratitude to our **Chairman Shri.P.SRIRAM** and all trust members of Chennai Institute of Technology for providing the facility and opportunity to do this project as a part of our undergraduate course.

We are grateful to our Principal **Dr.A.RAMESH M.E**, **Ph.D.** for providing us the facility and encouragement during the course of our work.

We sincerely thank our Head of the Department **Dr.R.Gowri** Department of CSE(Artificial Intelligence & Machine Learning) for having provided us valuable guidance, resources and timely suggestions throughout our work.

We would like to extend our thanks to our Project Co-ordinator **Dr.P.Karthikeyan** of the Department of CSE(Artificial Intelligence & Machine Learning), for his valuable suggestions throughout this project.

We wish to extend our sincere thanks to all Faculty members of the for their valuable suggestions and their kind cooperation for the successful completion of our project.

We wish to acknowledge the help received from the **Lab Instructors of the** Department of CSE (Artificial Intelligence & Machine Learning) and others for providing valuable suggestions and for the successful completion of the project.

NAME: THRISHA VIJAYAKUMAR **REG.NO:22AM059**

PREFACE

I, a student in the Department of CSE (Artificial Intelligence & Machine Learning) need to undertake a project to expand my knowledge. The main goal of my Core Course Project is to acquaint me with the practical application of the theoretical concepts I've learned during my course.

It was a valuable opportunity to closely compare theoretical concepts with real-world applications. This report may depict deficiencies on my part but still it is an account of my effort.

The results of my analysis are presented in the form of an industrial Project, and the report provides a detailed account of the sequence of these findings. This report is my Core Course Project, developed as part of my 3rd year project. As an engineer, it is my responsibility to contribute to society by applying my knowledge to create innovative solutions that address their changes.

1.ABSTRACT

This project focuses on the development of an AI-driven chatbot that interacts with PDF documents to extract and interpret information, providing users with a seamless question-and-answer experience. The chatbot leverages Generative AI technologies to understand natural language queries and generate relevant responses based on the content within the PDF files. The implementation is built using Google Colab, which serves as the development platform for handling PDF extraction and processing using Python, along with integration of OpenAI's API to harness advanced language understanding capabilities.

Streamlit is used to design an intuitive web interface that allows users to upload PDFs and engage with the chatbot in real-time, querying specific details or summarizing sections of the document. The integration of these tools results in a scalable, user-friendly solution that can be applied across various domains, such as education, research, and customer support, where interaction with extensive document datasets is essential.

The project report discusses the technical approach, including the preprocessing of PDFs, API integration, and web application development. It also highlights key challenges, such as handling diverse document formats and ensuring accurate responses, along with proposed solutions. Finally, the project offers insights into the deployment strategies on cloud platforms to enhance accessibility and scalability. This chatbot presents an efficient and effective way of automating document-related queries, reducing the need for manual search and reading processes.

2.INTRODUCTION

In today's information-driven world, PDFs are one of the most commonly used formats for sharing and storing documents. Whether in academia, business, or everyday life, these files often contain a wealth of information that users need to access quickly and efficiently. However, manually searching through extensive PDFs to extract relevant data can be time-consuming and cumbersome. To address this challenge, advancements in artificial intelligence (AI) and natural language processing (NLP) have paved the way for the creation of intelligent systems capable of automating the extraction and retrieval of information from such documents.

This project aims to develop a PDF chatbot using Generative AI, with the ability to comprehend, process, and respond to user queries about the content of a given PDF file. By integrating tools such as Google Colab for development, OpenAI's API for AI-driven language understanding, and Streamlit for building a user-friendly interface, the chatbot provides a scalable solution to interact with documents in a conversational manner. Users can upload PDFs, ask questions about their content, and receive relevant responses in real-time, greatly improving the efficiency of accessing information.

The primary objective of this project is to simplify document interaction by combining the capabilities of natural language models with practical web deployment solutions. This project report covers the methodology, from text extraction to AI integration, and the deployment of the system as a functional web application. The PDF chatbot can serve various use cases, including research, legal document analysis, and customer support, offering users a powerful tool to enhance productivity and streamline information retrieval.

3.LITERATURE SURVEY

- **Natural Language Processing (NLP):**
  - Traditional models like TF-IDF and LSA were used for text extraction.
  - Modern transformer models like BERT and GPT-3 revolutionized NLP by understanding and generating natural language.
  - GPT-3, used in this project, is highly effective for handling conversational queries about PDF content.
- **Generative AI for Chatbots:**
  - Generative models, particularly GPT-3, enable dynamic and contextually accurate chatbot interactions.
  - These AI chatbots surpass rule-based systems in handling complex queries and retrieving relevant document information.
- **PDF Extraction Techniques:**
  - Tools like PyPDF2, PDFMiner, and PyMuPDF are widely used for extracting text from PDFs.
  - These libraries handle various document structures and ensure that the chatbot processes content accurately.
- **Web App Deployment with Streamlit:**
  - Streamlit simplifies the deployment of AI applications, offering an interactive, user-friendly web interface.
  - It allows users to upload PDFs, ask queries, and receive responses in real-time without extensive frontend development.

4.Project Objective:

The objective of this project is to design and develop an AI-powered chatbot capable of interacting with PDF documents to enhance information accessibility and retrieval. This chatbot will leverage cutting-edge Generative AI, specifically OpenAI's language models, along with natural language processing (NLP) and PDF extraction techniques to enable intelligent, context-aware conversations. The key goals of this project include:

- **Automated PDF Querying:** Develop a system that allows users to upload PDF documents and ask natural language questions related to their content, simplifying the process of retrieving specific information.
- **Accurate and Contextual Responses:** Ensure the chatbot provides accurate and contextually relevant responses by using advanced AI models like GPT-3, capable of understanding user queries and comprehending the content within the document.
- **Real-Time Interaction:** Implement a user-friendly web interface using Streamlit, allowing users to interact with the chatbot in real-time. The interface will support seamless PDF uploads and enable dynamic, conversational exchanges with the document.
- **Efficient Document Processing:** Use Python libraries such as PyPDF2 and PDFMiner to extract and process text from PDFs effectively, ensuring accurate handling of various document structures, layouts, and elements.
- **Scalability and Cloud Integration:** Deploy the solution on a cloud-based environment, utilizing platforms like Google Colab for model training and execution, ensuring the system is scalable and accessible to a wide range of users.

PROJECT SCOPE

This project focuses on developing an AI-powered PDF chatbot that enables users to interact with and extract information from PDF documents through natural language queries. The scope includes implementing text extraction techniques to handle diverse document formats and leveraging Generative AI models for accurate response generation. The chatbot will be deployed using Streamlit, providing a user-friendly web interface for real-time interaction.

5.TECHSTACK USED

Python: The backbone of the development process, Python is employed for implementing core functionalities, including AI model integration, PDF processing, and user interaction.

OpenAI GPT-3: This advanced generative model is used for understanding user queries and generating contextually relevant responses based on the content of the uploaded PDF documents. Its ability to interpret natural language makes it a perfect fit for the chatbot's conversational capabilities.

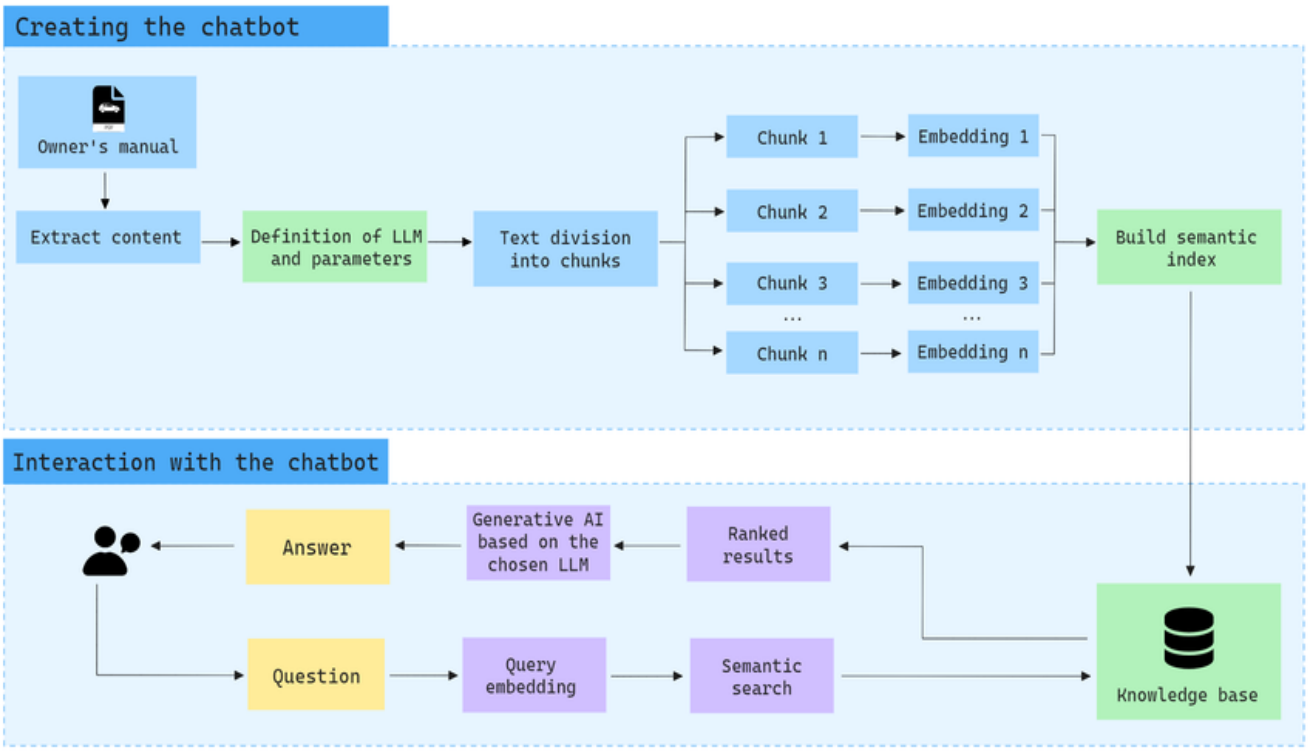
PyPDF2: A Python library that facilitates the extraction of text and metadata from PDF files, allowing the chatbot to retrieve relevant content.

PDFMiner: Used for more complex text extraction tasks, PDFMiner provides detailed access to the structure and layout of PDFs, ensuring accurate data retrieval from various document formats.

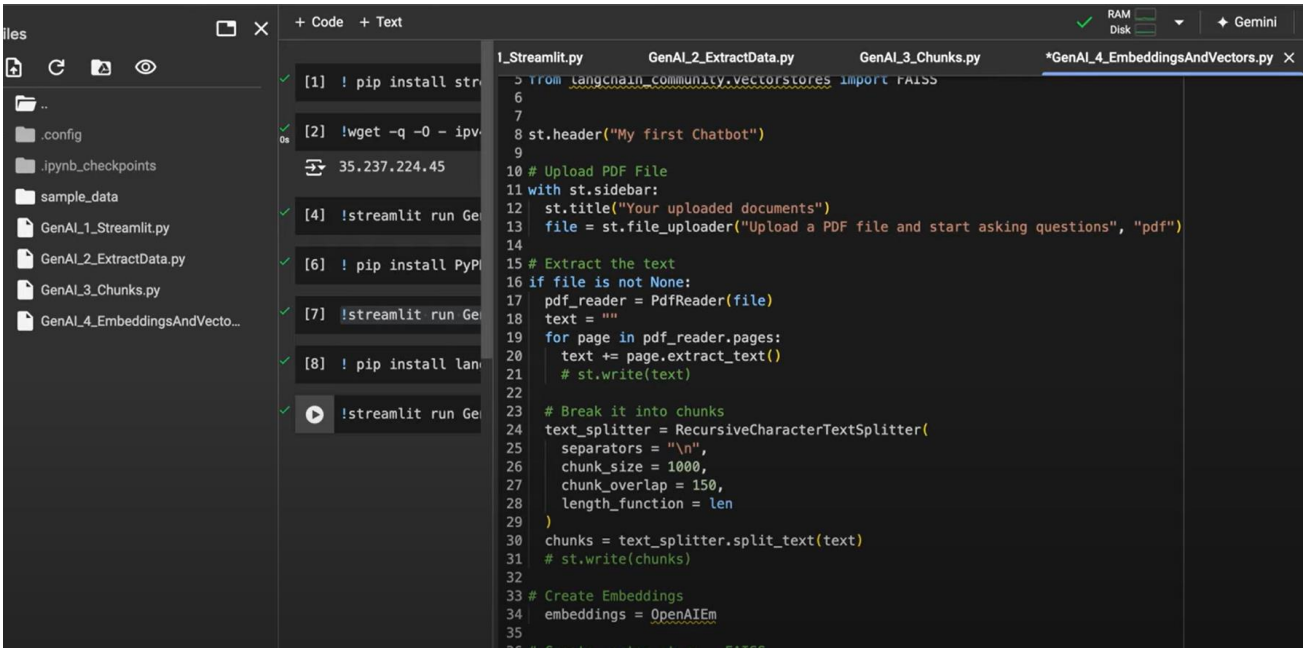
Streamlit: This framework allows for the rapid development of interactive web applications. Streamlit provides a user-friendly interface where users can upload PDFs and interact with the chatbot in real-time, enhancing overall user experience.

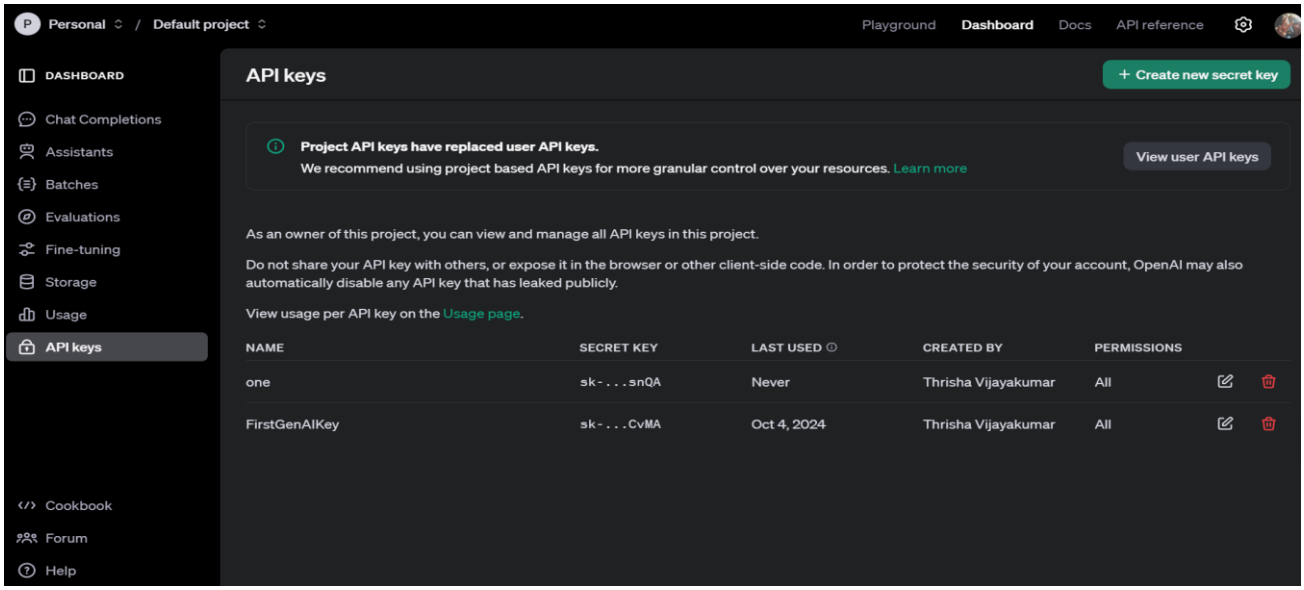
Google Colab: An online platform that supports collaborative coding and offers powerful GPU resources for training and deploying machine learning models

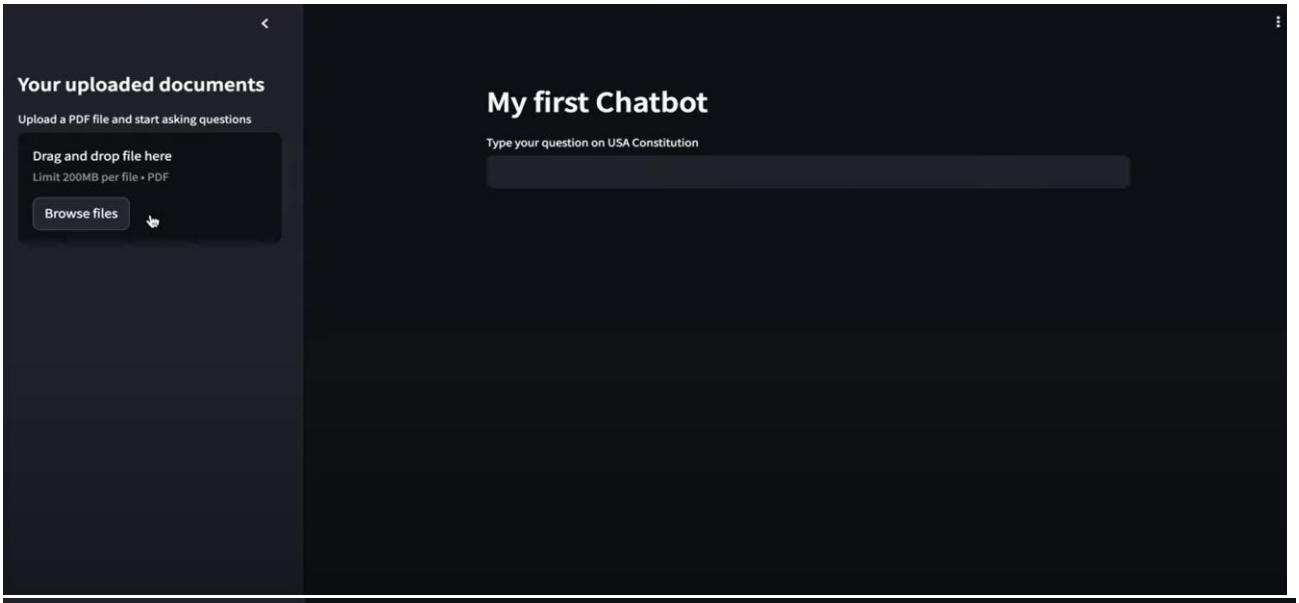
6.ARCHITECTURE DIAGRAM

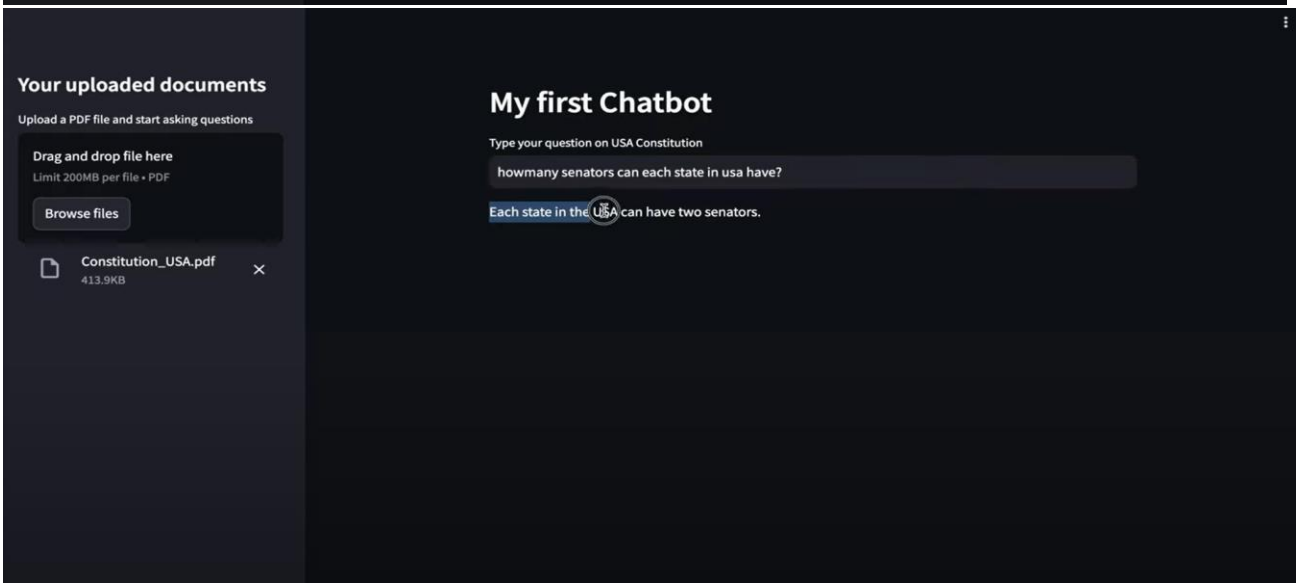


7.IMPLEMENTATION AND PROCEDURE




Personal / Default project
Playground
Dashboard
Docs
API reference
DASHBOARD
Chat Completions
Assistants
Batches
Evaluations
Fine-tuning
Storage
Usage
API keys
Cookbook
Forum
Help
API keys
+ Create new secret key
Project API keys have replaced user API keys.
We recommend using project based API keys for more granular control over your resources. Learn more
View user API keys
As an owner of this project, you can view and manage all API keys in this project.
Do not share your API key with others, or expose it in the browser or other client-side code. In order to protect the security of your account, OpenAI may also automatically disable any API key that has leaked publicly.
View usage per API key on the Usage page.
NAME
SECRET KEY
LAST USED
CREATED BY
PERMISSIONS
one
sk-...snQA
Never
Thrisha Vijayakumar
All
FirstGenAIKey
sk-...CvMA
Oct 4, 2024
Thrisha Vijayakumar
All


Your uploaded documents
Upload a PDF file and start asking questions
Drag and drop file here
Limit 200MB per file • PDF
Browse files
My first Chatbot
Type your question on USA Constitution


Your uploaded documents
Upload a PDF file and start asking questions
Drag and drop file here
Limit 200MB per file • PDF
Browse files
Constitution_USA.pdf
413.9KB
My first Chatbot
Type your question on USA Constitution
howmany senators can each state in usa have?
Each state in the USA can have two senators.

9.REFERENCES

- Jones, K. S., "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, 1972.
- Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- Brown, T. B., et al., "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*, 2020.
- Radford, A., et al., "Improving Language Understanding by Generative Pre-Training," *OpenAI*, 2018.
- Bommasani, R., et al., "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258*, 2021.
- Dale, R., "The return of the chatbots," *Natural Language Engineering*, 2016.
- Liu, Y., et al., "DocBot: An AI-powered Conversational Agent for Document Analysis," *IEEE Access*, 2020.
- Schafer, D., "Extracting Text from PDFs with Python," *Real Python*, 2018.
- Clarke, C., et al., "Text Mining for Information Extraction from PDF," *Journal of Information Science*, 2019.
- Batts, J., "Rapid Prototyping of Web Apps using Streamlit," *Journal of Web Development*, 2020.