Project Title

PolicyBot – Internal Policy Document Assistant

■ Problem Statement

Large organizations often have thousands of internal policies, guidelines, compliance documents, and SOPs spread across multiple PDFs or internal wikis. Employees and managers struggle to quickly find relevant and up-to-date information when they have questions about leave policies, expense claims, IT support, or HR processes.

■ Goal

To build an LLM-powered semantic search chatbot using LangChain or LlamaIndex, capable of answering employee questions based on internal policy documents (PDFs, Word docs, or plain text), with cited references and extracted context.

■ Tools & Stack

- LangChain or LlamaIndex (Core framework)
- OpenAI or Gemini (LLM API for answering queries)
- ChromaDB or FAISS (Vector DB for storing document chunks)
- Streamlit or Gradio (for a web-based chat UI)
- PyMuPDF / PDFplumber (for PDF reading and chunking)
- Tiktoken / Sentence Transformers (for efficient chunking + embedding)

■ Sample Data

- Public HR policy PDFs (Google “Company HR policies filetype:pdf”)
- Kaggle: HR Policies Dataset
- Custom-made SOPs or university handbooks

■ Key Features

- Upload PDF/Word documents or load from directory
- Document chunking, embedding, and indexing using LangChain or LlamaIndex
- Query interface for employees to ask questions in natural language
- LLM retrieves context + generates answer
- Shows exact policy section used (with reference)
- Option to cite source, copy answer, or download as PDF

■ Bonus Enhancements

- Memory: Save past chat history with retrieval
- Multi-file Support: Allow uploading multiple PDFs simultaneously
- Access Controls: Restrict sections based on user role (Manager, Employee, HR)
- Offline Mode: Run on local models like Mistral/LLama3 with LangChain

■ Deliverables

- Streamlit-based web application
- Modular Python code with utils, config, and tests
- .env file for API key setup
- Full project ZIP with data, index, app, and documentation
- Project logo and UI design screenshot
- This documentation PDF

■ Summary for Submission

This project fits within the "Search Systems" domain using LangChain or LlamaIndex, demonstrates end-to-end LLM-based application development, and tackles a real-world business challenge in internal knowledge access.

■ Comprehensive Documentation

Project Goals:

Build an intelligent assistant that enables employees to ask questions about organizational policies and retrieve answers from official documents instantly. Ensure transparency, accuracy, and usability.

Data Sources:

- Internal PDF and Word documents related to HR, compliance, IT, etc.

- Public datasets like Kaggle’s HR Policies

- User-uploaded documents

Design Choices:

- Streamlit for a lightweight, interactive UI

- LangChain for chaining LLMs and document retrievers

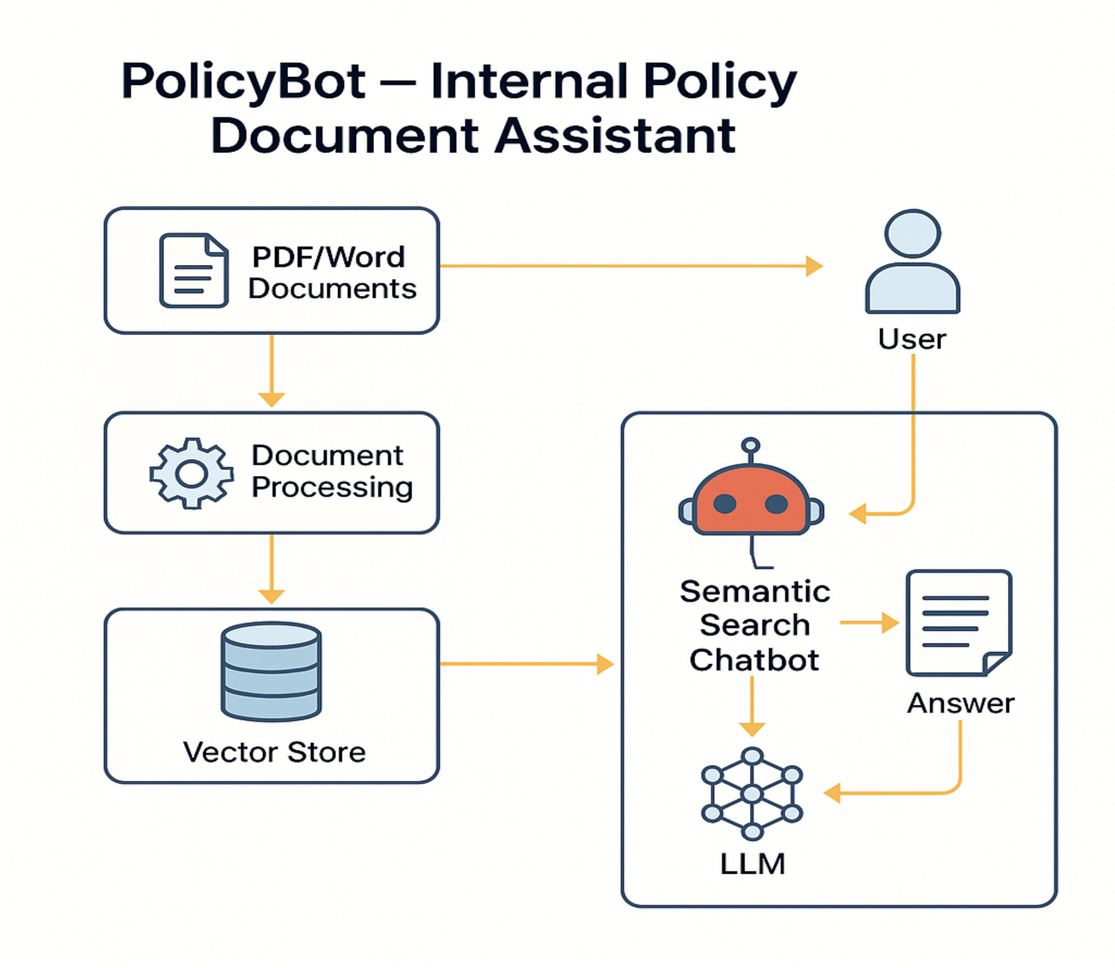
- ChromaDB for persistent vector storage

- .env file for key security

Challenges Faced:

- Handling various document formats and text extraction
- Ensuring embeddings are context-aware
- Managing performance and cost of LLM calls
- Providing citations to improve answer credibility

■ System Design Flowchart


PolicyBot – Internal Policy Document Assistant
PDF/Word Documents
User
Document Processing
Semantic Search Chatbot
Answer
Vector Store
LLM