MEET THE TEAM

Akshita

Tanisha

Ajinkya
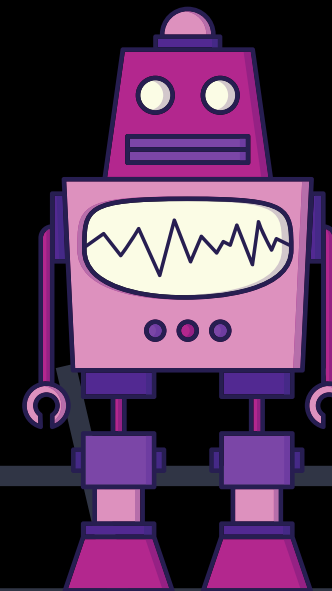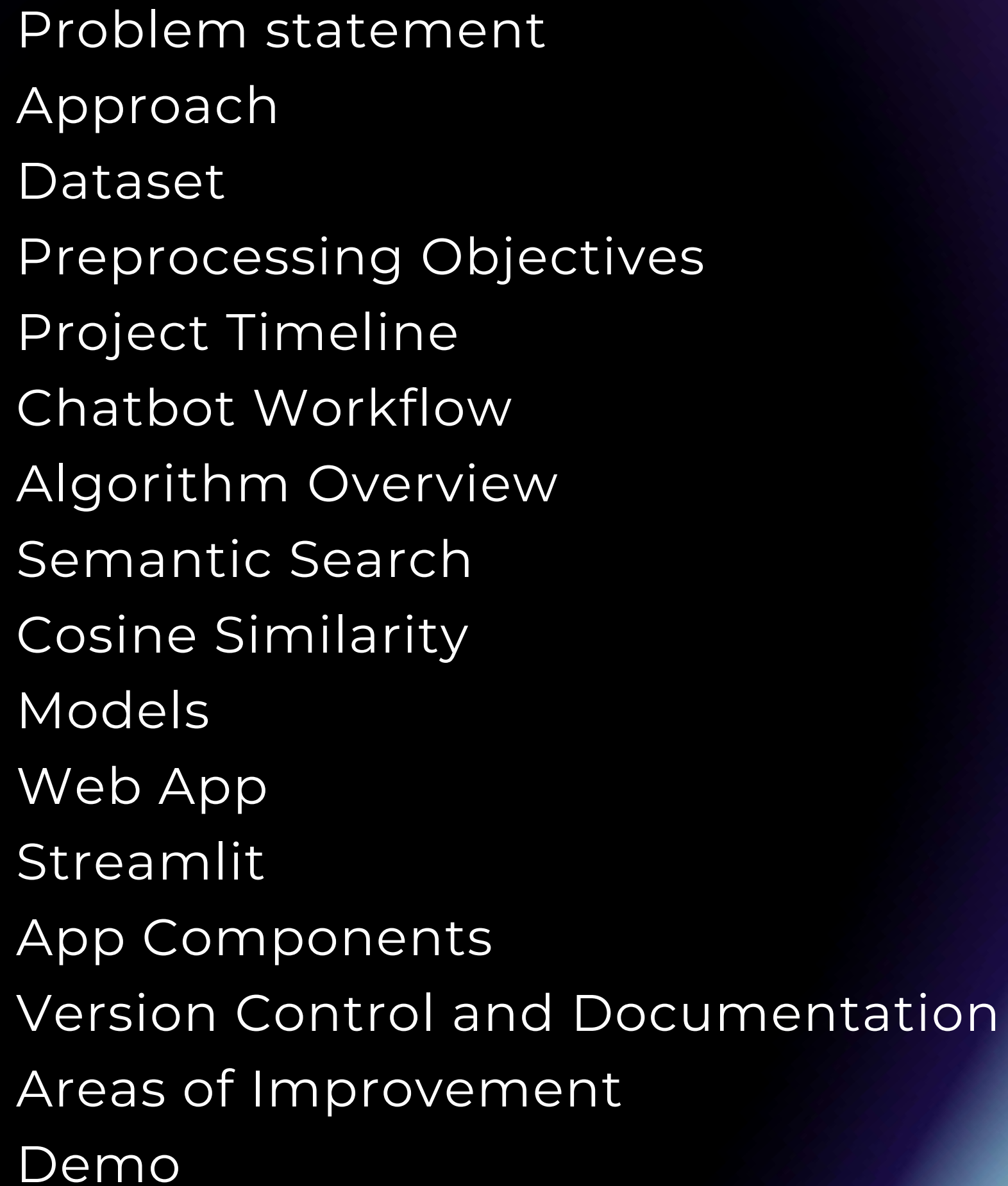
Wissame

Swachhith

CONTENT

- Problem statement
- Approach
- Dataset
- Preprocessing Objectives
- Project Timeline
- Chatbot Workflow
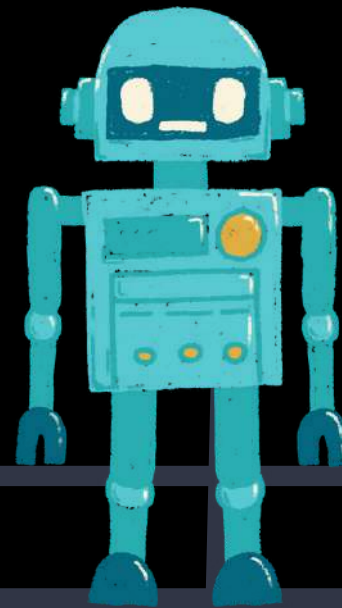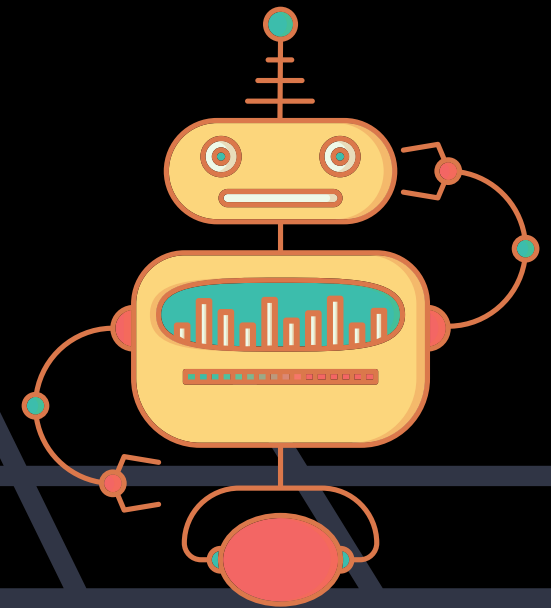- Algorithm Overview
- Semantic Search
- Cosine Similarity
- Models
- Web App
- Streamlit
- App Components
- Version Control and Documentation
- Areas of Improvement
- Demo

PROBLEM STATEMENT

To develop and deploy a chatbot to assist the Legal and IP team of MANE.

user has sent the following prompt: yes

When is a CDA needed ?

When Mane will share confidential information with another person or entity. Also, when someone will observe or learn about Mane's confidential information. Additionally, when any of the parties (MANE and Counterparty) wish to disclose confidential information to each other.

Is this the information you were looking for ?

yes

APPROACH

Understanding Needs

Data Collection & Preprocessing

Model Selection

Data Augmentation

Model Training and Evaluation

Streamlit Integration

Deployment

Documentation

DATASET

Total number of files: 79

TYPES OF FILES

01 — PDF

02 — DOC

03 — DOCX

04 — PPT

PREPROCESSING OBJECTIVES
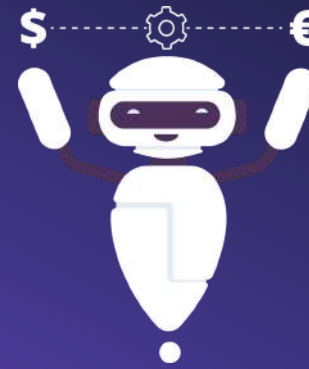
CONVERSION TO TEXT FILES

Extracting content from various formats like PDF, DOC, and DOCX, with additional steps such as lemmatization to normalize words. This prepares the data for natural language processing tasks by creating a standardized and linguistically processed text.

DATA AUGMENTATION

We had less and unstructured data

STRUCTURING

- Filename
- Context
- Question
- Answer

Project Timeline

WEEK 1

Initilization and Data Understanding

WEEK 2

Chosing LLM and Data Conversion

WEEK 3

Initial LLM Model Training and Challenges

WEEK 4

Transition to BERT

WEEK 5

BERT vs SBERT and Front End Development

WEEK 6

Mutual Data Creation and Organized Output

WEEK 7

Finetuning and Integration
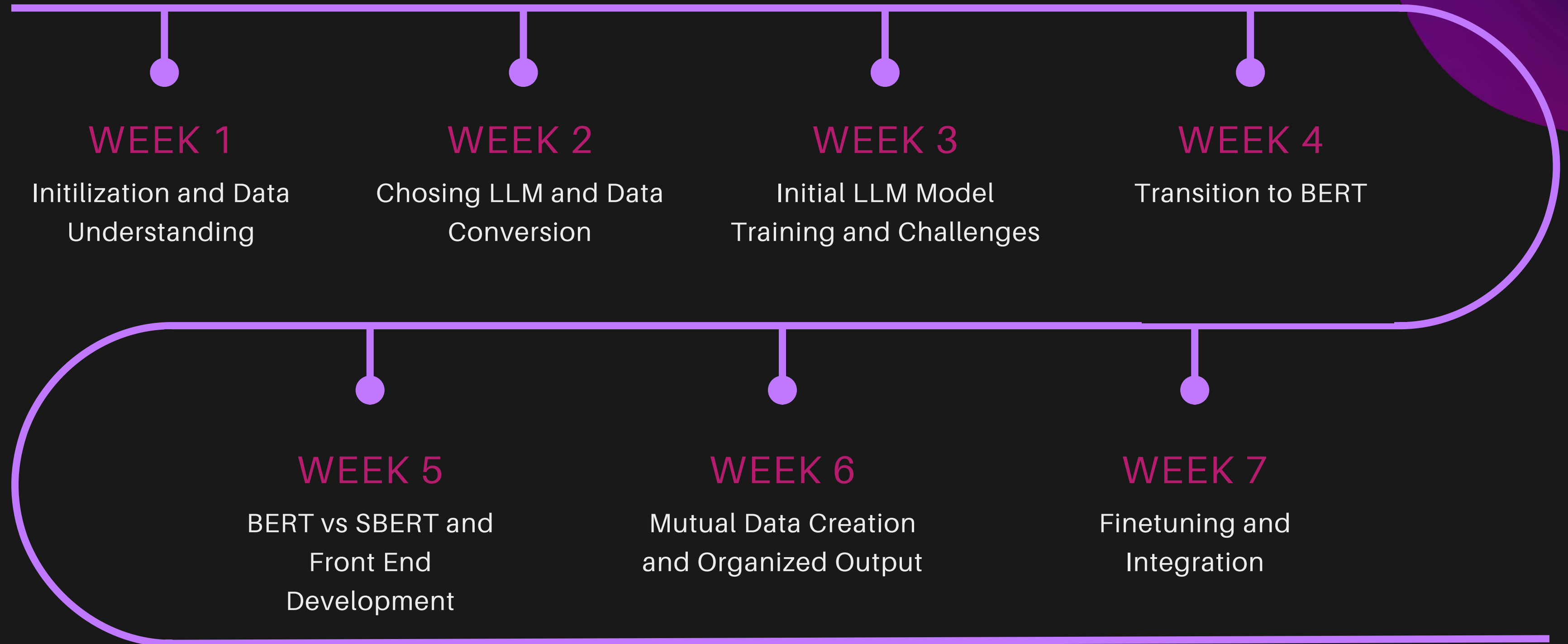
CHATBOT WORKFLOW AND ALGORITHM OVERVIEW

Workflow

Source —— PyMuPdf, Python-Docx

Selected k chunk

Text Extraction → **Conversion to chunks** (Spacy, NLTK) → **Similarity Search** → **Prompt template+ questions+chunk sizes**

Queries and response from user

Streamlit App → **Response from model & Accuracy**

Vicuna
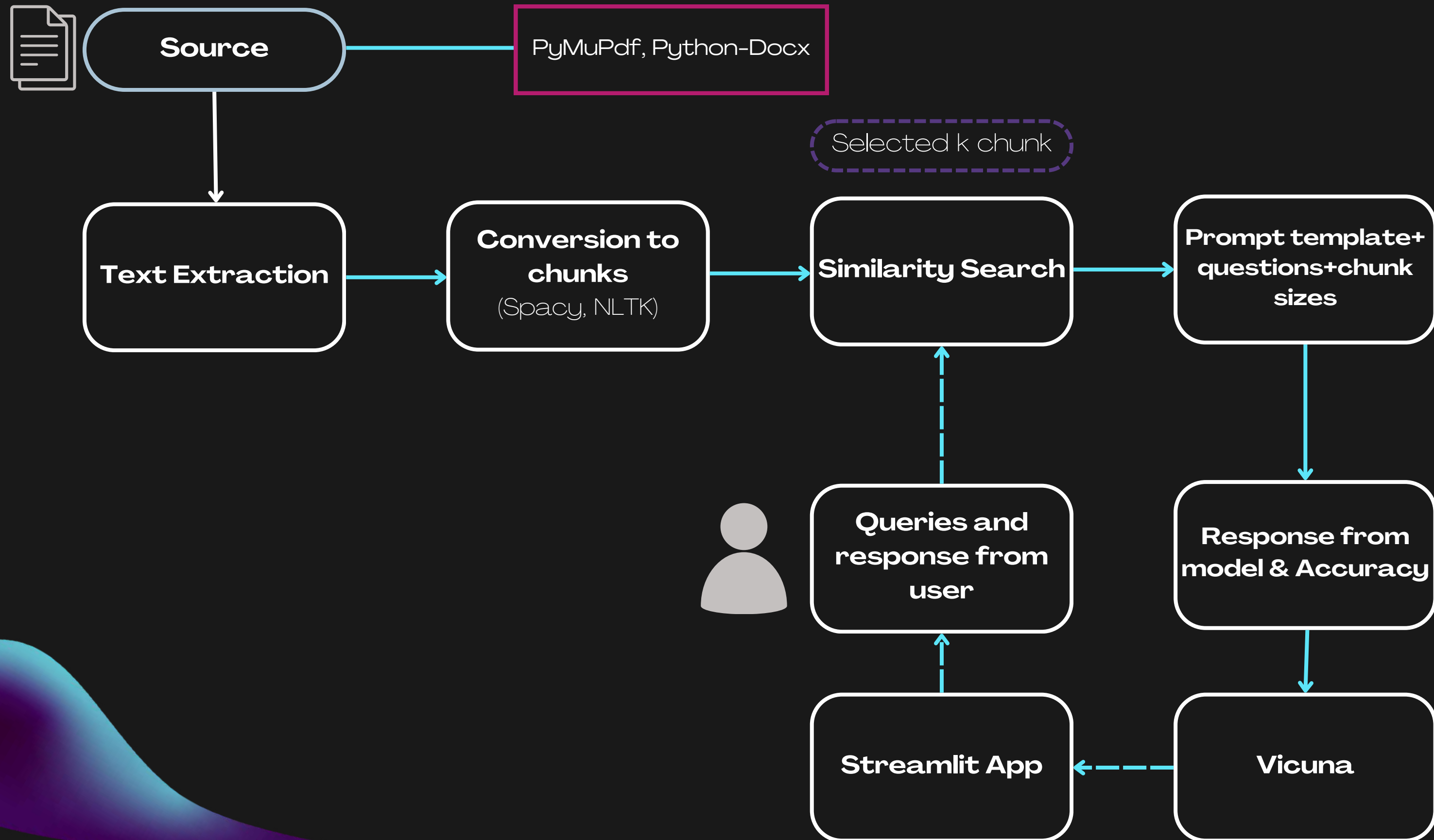
1. Initialization
- Model and Data Loading
- User Interface (UI) Setup
2. User Interaction
- User Input Handling
- Similarity Computation
- Displaying the Response
3. Thresholds for Response
4. User satisfaction check
- Satisfaction Handeling
- Medium Similarity
- Low Similarity

ALGORITHM OVERVIEW

SEMANTIC SEARCH

Necessity

Match making beyond traditional methods

Capturing Context

Consider meaning or semantics of words/phrases

Adaptability

Handling wide range of user inputs
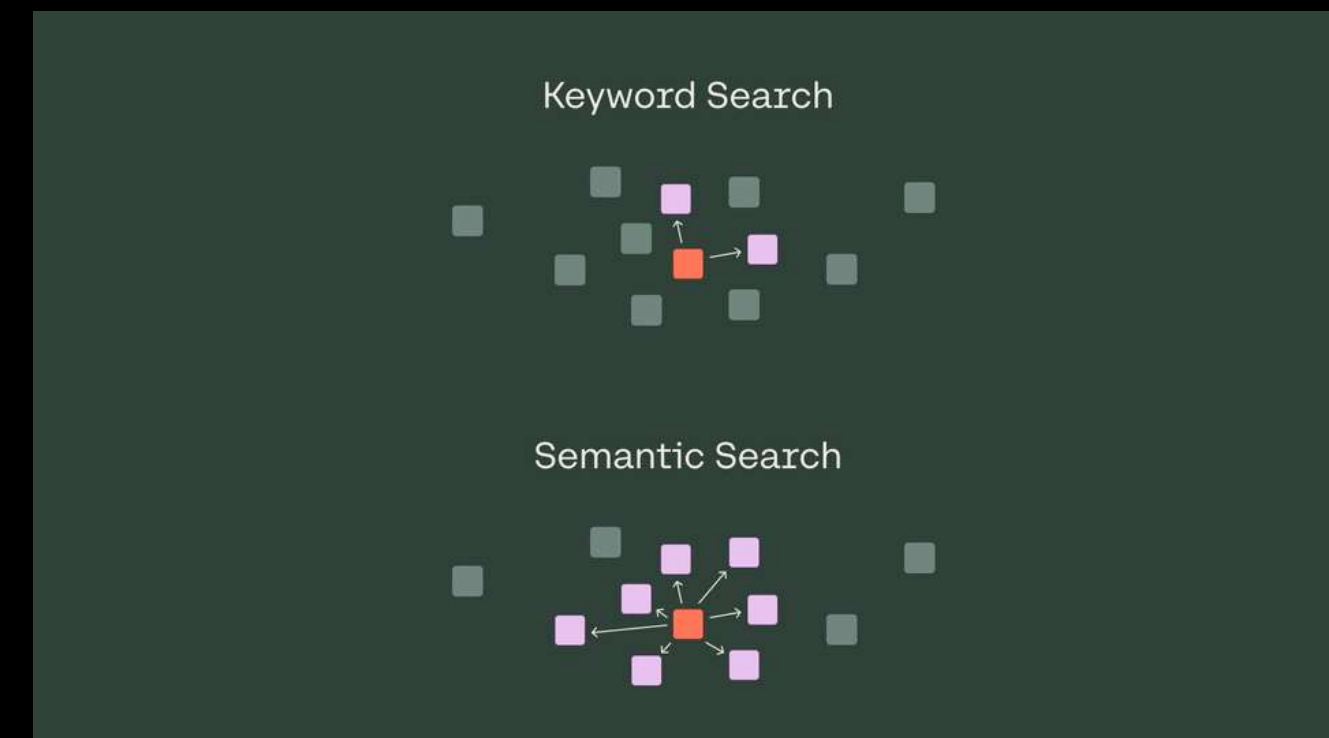
COSINE SIMILARITY

Matching Vectors

Identifying similar context vectors

Effective Retrieval

Matching without exact words in sequence

- Angle θ close to 0
- Cos(θ) close to 1
- **Similar vectors**

- Angle θ close to 90
- Cos(θ) close to 0
- **Orthogonal vectors**

- Angle θ close to 180
- Cos(θ) close to -1
- **Opposite vectors**

Models

ALL-MPNET-BASE-V2

MULTI-QA-MPNET-BASE-DOT-V1

ALL-DISTILROBERTA-V1

MULTI-QA-MINILM-L6-COS-V1
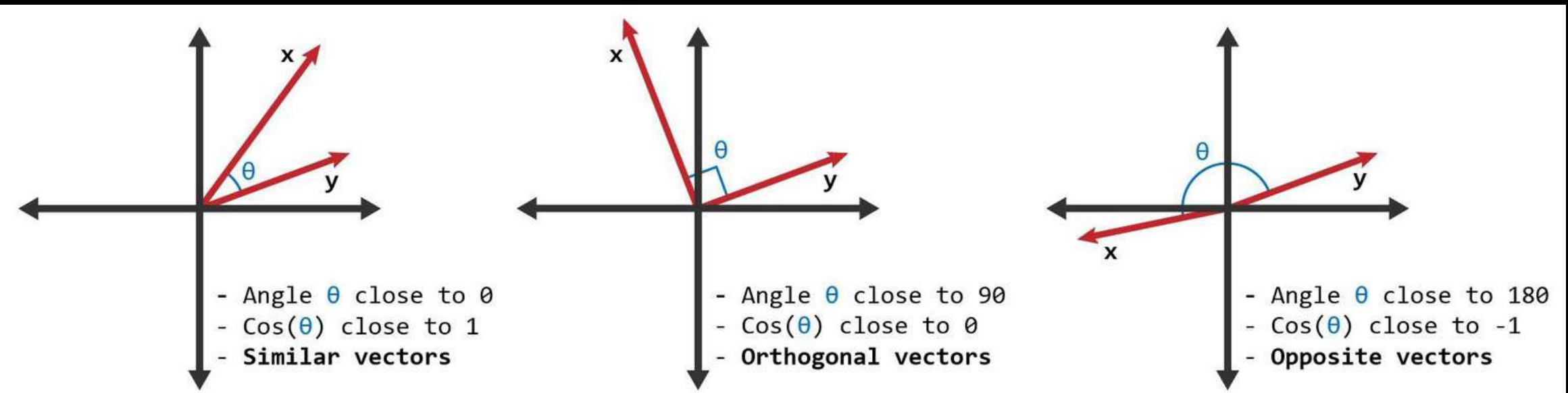
PARAPHRASE-MINILM-L3-V2

Ref:
https://www.sbert.net/docs/pretrained_models.html

MODELS COMPARISON

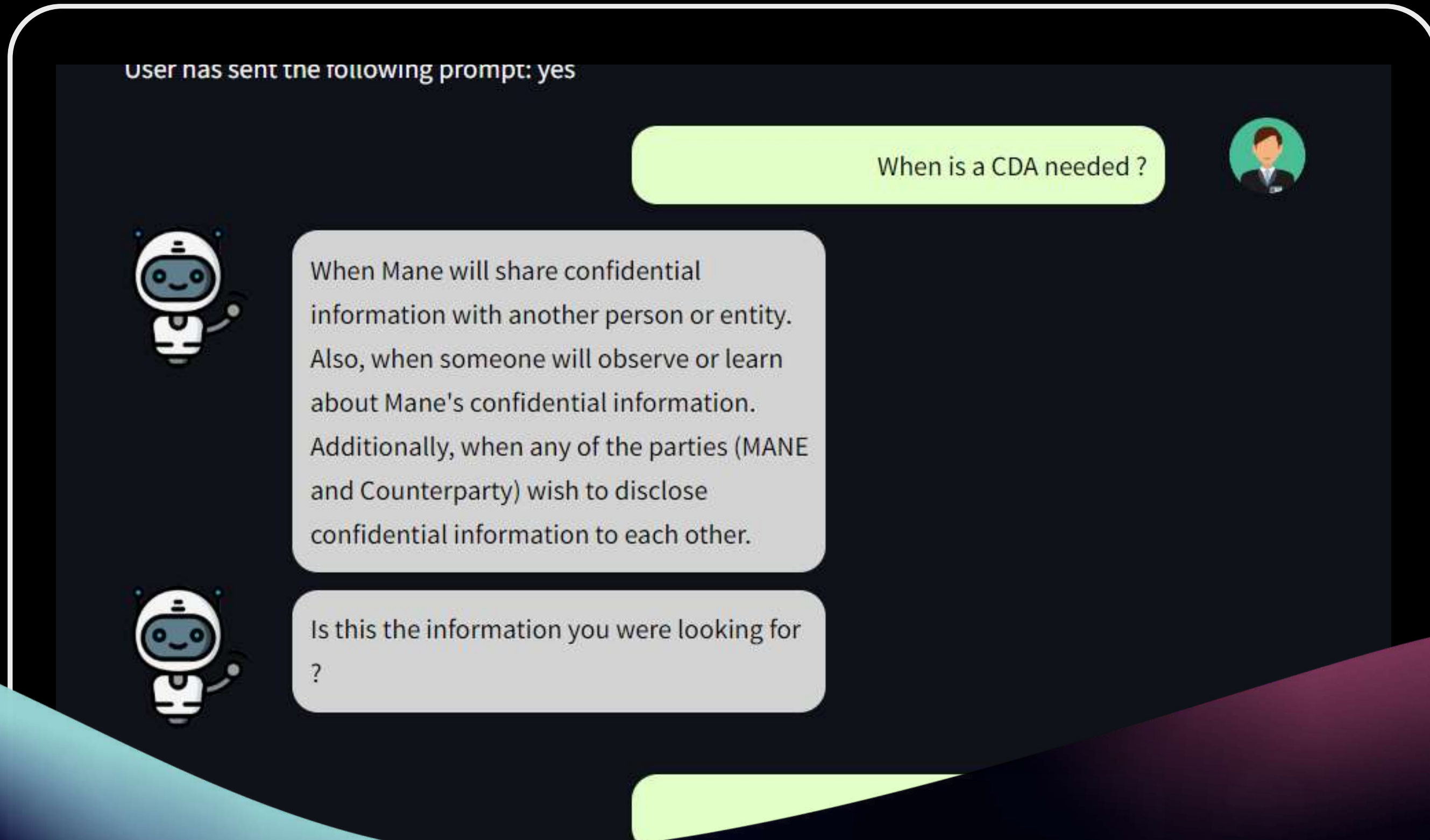| User Question	Retrieved relevant question	model's name	cosine similarity
Briefly explain what trade secrets mean.	What is a Trade Secret?	multi-qa-mpnet-base-dot-v1	0.8985
Briefly explain what trade secrets mean.	What is a Trade Secret?	all-distilroberta-v1	0.8452
Briefly explain what trade secrets mean.	What is a Trade Secret?	all-mpnet-base-v2	0.8883
Briefly explain what trade secrets mean.	What is a Trade Secret?	multi-qa-MiniLM-L6-cos-v1	0.9427
Briefly explain what trade secrets mean.	What constitutes a trade secret?	paraphrase-MiniLM-L3-v2	0.8603
What are the suggestions for an agent acting v	What power does the President of the Office have re	multi-qa-mpnet-base-dot-v1	0.6023
What are the suggestions for an agent acting v	what measures can a supervisory authority undertal	all-distilroberta-v1	0.5857
What are the suggestions for an agent acting v	What authority does the President of the Office have	all-mpnet-base-v2	0.5907
What are the suggestions for an agent acting v	What are the suggestions for an agent acting withou	multi-qa-MiniLM-L6-cos-v1	0.5243
What are the suggestions for an agent acting v	what measures can a supervisory authority undertal	paraphrase-MiniLM-L3-v2	0.5266
What are the key actions in crisis managemen	What are the primary actions that the DPO and the	multi-qa-mpnet-base-dot-v1	0.7324
What are the key actions in crisis managemen	What are the primary actions that the DPO and the	all-distilroberta-v1	0.6743
What are the key actions in crisis managemen	What are the primary actions that the DPO and the	all-mpnet-base-v2	0.6788
What are the key actions in crisis managemen	What are the primary actions that the DPO and the	multi-qa-MiniLM-L6-cos-v1	0.6159
What are the key actions in crisis managemen	What does the policy say about the consequences o	paraphrase-MiniLM-L3-v2	0.5632
What are the duties of the author and witness	What is the role of the witness in the laboratory not	multi-qa-mpnet-base-dot-v1	0.7548
What are the duties of the author and witness	What are the duties of the witness regarding the lab	all-distilroberta-v1	0.5792
What are the duties of the author and witness	What is the role of the witness in the laboratory not	all-mpnet-base-v2	0.7376
What are the duties of the author and witness	What is the role of the witness in the laboratory not	multi-qa-MiniLM-L6-cos-v1	0.5968
What are the duties of the author and witness	What are the duties of the witness regarding the lab	paraphrase-MiniLM-L3-v2	0.5761

MODEL COMPARISON (GRAPHICAL)

- multi-qa-mpnet-base-dot-v1
- all-distilroberta-v1
- all-mpnet-base-v2
- multi-qa-MiniLM-L6-cos-v1
- paraphrase-MiniLM-L3-v2
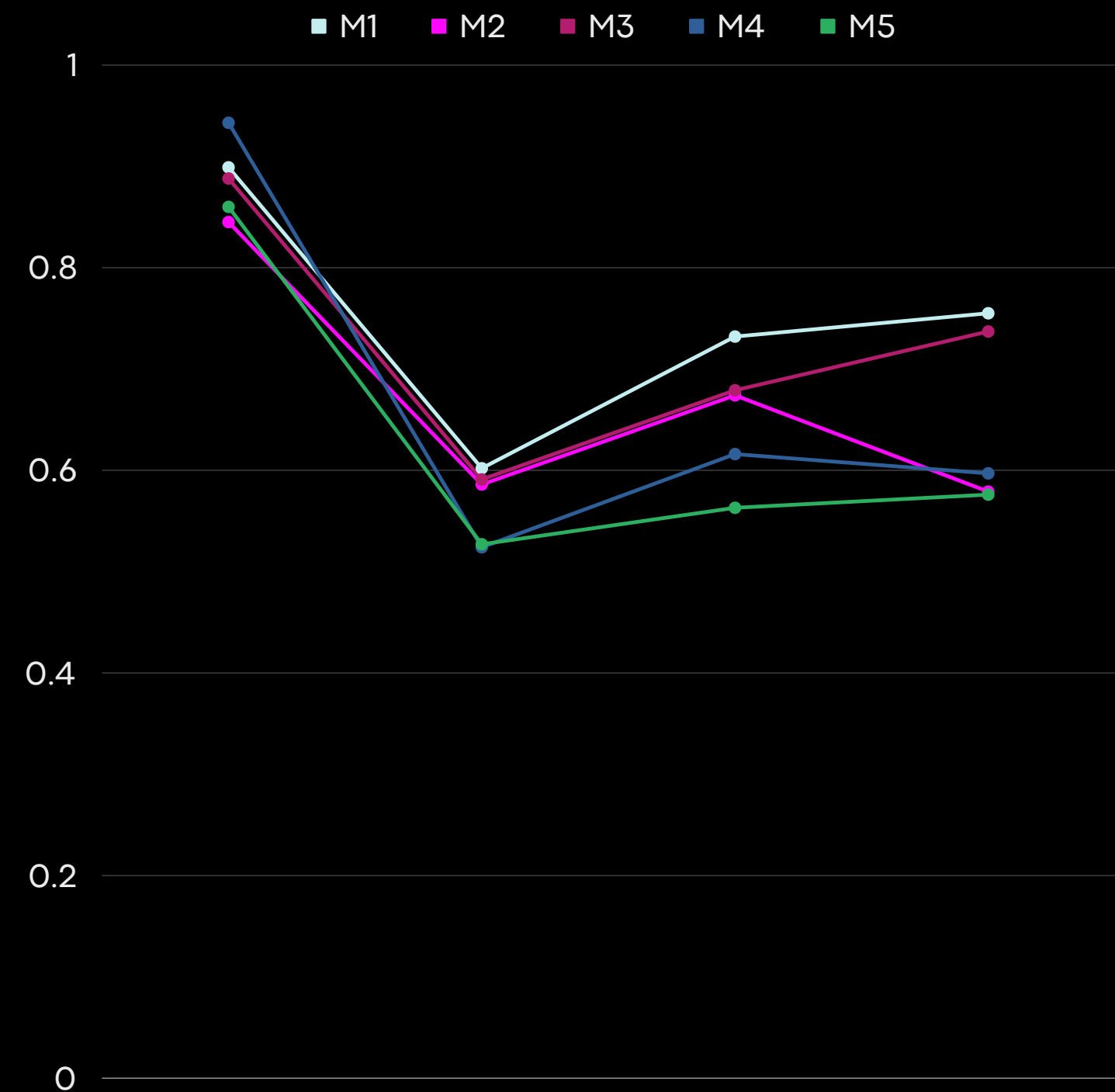
WEB APP

Interactive Interface:

User has sent the following prompt: yes

When is a CDA needed ?

When Mane will share confidential information with another person or entity. Also, when someone will observe or learn about Mane's confidential information. Additionally, when any of the parties (MANE and Counterparty) wish to disclose confidential information to each other.

Is this the information you were looking for ?

Streamlit

Open source python framework of ML and Data Science

1 Well-defined functions for each action and Streamlit UI

2 Easy to collaborate with our sbert model

3 Simple to launch using localhost ports
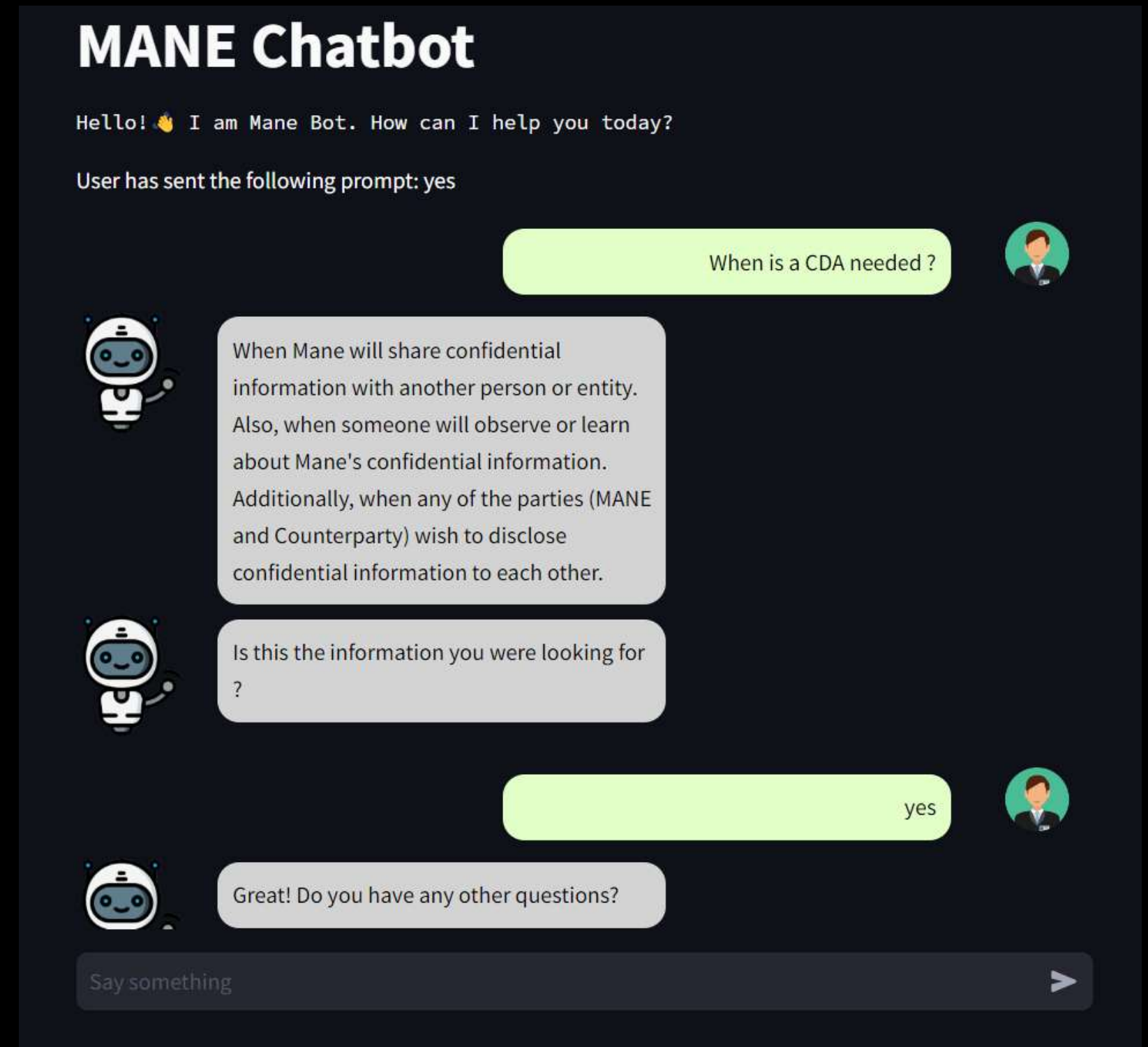
4 Ease of maintenance and modification


```
Mane-Project > Model > ⌘ draft2copy.py > ⋃ handle_details > [∅] most_similar_index
117                 st.image(chatbot_image, width=40, output_format="png", use_column_width=True)
118             with col2:
119                 st.markdown(f"<div class='chat-bubble bot-bubble'>{message}</div>", unsafe_allow_html=True)
120
121     def display_chat_history():
122         for role, message in st.session_state.chat_history_list:
123             if role == "USER":
124                 st.image(USER_image, width=40, output_format="png", use_column_width=True)
125                 st.markdown(f"<div class='chat-bubble USER-bubble'>{message}</div>", unsafe_allow_html=True)
126             else:
127                 st.image(chatbot_image, width=40, output_format="png", use_column_width=True)
128                 st.markdown(f"<div class='chat-bubble bot-bubble'>{message}</div>", unsafe_allow_html=True)
129
130     # Main interaction loop
131     def handle_interaction(prompt):
132         if st.session_state.conversation_state == CONVERSATION_STATES['INITIAL']:
133             process_initial_prompt(prompt)
134         elif st.session_state.conversation_state == CONVERSATION_STATES['AWAITING_CONFIRMATION']:
135             handle_confirmation(prompt)
136         elif st.session_state.conversation_state == CONVERSATION_STATES['AWAITING_DETAILS']:
137             handle_details(prompt)
138         elif st.session_state.conversation_state == CONVERSATION_STATES['AWAITING_NUMERIC_INPUT']:
139             if prompt.isdigit():
140                 handle_numeric_input(prompt)
141             else:
142                 add_to_chat_history("ASSISTANT", "That's not a number. Please enter the number corresponding to your question.")
143
144
145     def process_initial_prompt(prompt):
146         if 'last_answered_filename' in st.session_state and prompt.lower().strip() in ["what is the reference for the last answer?", "source of last a
147             if st.session_state.last_answered_filename:
148                 add_to_chat_history("ASSISTANT", f"The reference for the last answer is: {st.session_state.last_answered_filename}")
149             else:
150                 add_to_chat_history("ASSISTANT", "I don't have a reference for the last answer.")
151         else :
152             USER_input_embedding = model.encode(preprocess_text(prompt))
153             similarities = util.cos_sim(USER_input_embedding, combined_embeddings)[0]
```
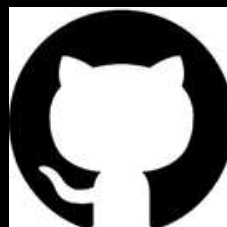
App Components

- USER AND BOT INTERCATION

- VICUNA INTEGRATION

- FILENAME IDENTIFIER
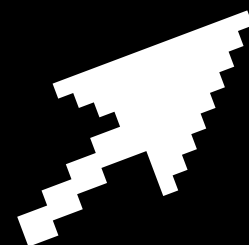
- MINIMALIST DESIGN

- USER-FRIENDLY INTERFACE

MANE Chatbot

Hello! 👋 I am Mane Bot. How can I help you today?

User has sent the following prompt: yes

When is a CDA needed ?

When Mane will share confidential information with another person or entity. Also, when someone will observe or learn about Mane's confidential information. Additionally, when any of the parties (MANE and Counterparty) wish to disclose confidential information to each other.

Is this the information you were looking for ?

yes

Great! Do you have any other questions?

Say something

AREAS OF IMPROVEMENT

- Enhanced Natural Language Understanding
- Expansion of Document Coverage
- File retrieval (Not filename!)
- Integration with Additional Systems
- Personalization and User Experience
- Improved Feedback Mechanisms
- Enhanced Security and Compliance Measures

DEMO

MANE Chatbot

Hello! 👏 I am Mane Bot. How can I help you today?

User has sent the following prompt: Does a patent guarantee the right to use an invention?

Say something

THANK YOU!

Open to Q&A