Cognitive Computing on Personalized Data

Submitted in partial fulfillment of the requirements
of the degree of

Bachelor of Engineering in Information Technology

By

Ayush Tripathi (20101B0019)
Hasnain Sayyed (21101B2003)
Bhargav Mahajan (20101B0013)
Darshan Mahankar (20101B0002)

Under the Guidance of

Dr. Dilip Motwani
Department of Information Technology Engineering

(An Autonomous Institute Permanently Affiliated to Mumbai University)

Vidyalankar Institute of Technology

Wadala(E), Mumbai-400437

University of Mumbai

2023-24

CERTIFICATE OF APPROVAL

This is to certify that the project entitled

Cognitive Computing on Personalized Data

is a bonafide work of

Ayush Tripathi (20101B0019)

Hasnain Sayyed (21101B2003)

Bhargav Mahajan (20101B0013)

Darshan Mahankar (20101B0002)

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the

degree of

Undergraduate in **"INFORMATION TECHNOLOGY".**

Guide	Head of Department	Principal
Dr. Dilip Motwani	Dr. Vipul Dalal	Dr. S. A. Patekar

Project Report Approval for B. E.

This project report entitled ***Cognitive Computing on Personalized Data*** by

1. ***Ayush Tripathi (20101B0019)***
2. ***Hasnain Sayyed (21101B2003)***
3. ***Bhargav Mahajan (20101B0013)***
4. ***Darshan Mahankar (20101B0002)***

is approved for the degree of ***Bachelor of Engineering in Information Technology***

Examiners

1._______________________________

2._______________________________

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will cause disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of student	Roll No.	Signature
1. Ayush Tripathi	20101B0019	
2. Hasnain Sayyed	21101B2003	
3. Bhargav Mahajan	20101B0013	
4. Darshan Mahankar	20101B0002	

Date:

Acknowledgement

We are honored to present –Cognitive Computing on Personalized Data as our B.E Final Year Project. We are using this opportunity to express our profound gratitude to our principal –Dr. Sunil Patekar‖ for providing us with all proper facilities and support.

We express our deepest gratitude towards our Head of the Department –Dr. Vipul Dalal‖ for his valuable and timely advice during the various phases in our project. We would like to thank our project guide –Dr. Dilip Motwani for support, patience, and faith in our capabilities and for giving us flexibility in terms of working and reporting schedules. Finally, we would like to thank everyone who has helped us directly or indirectly in our project.

We would also like to thank our staff members and lab assistant for permitting us to use computers in the lab when required. We thank our college for providing us with excellent facilities that helped us to complete and present this project.

Ayush Tripathi
Hasnain Sayyed
Bhargav Mahajan
Darshan Mahankar

ABSTRACT

This project aims to revolutionize the interaction between users and Large Language Models (LLMs) by addressing the inherent limitation of short context length. By developing a web-based portal, we propose an innovative solution that seamlessly integrates custom data prompts with LLMs to enable more robust and contextually relevant responses. Unlike conventional approaches that rely solely on textual input, our portal supports a wide range of data types including text, images, audio, video, and YouTube videos. This versatility ensures that users can leverage diverse datasets to enrich the context for LLMs, thereby enabling more sophisticated understanding and generation of responses.

Through the portal, users gain the ability to engage in conversational interactions with their data without being encumbered by redundant or extraneous information. By inputting tailored queries or prompts, users can extract precise insights and information from their datasets. For instance, users in the medical field can upload patient records or case studies and pose specific questions related to diagnoses, treatments, or outcomes. The LLM, augmented by the contextual information provided by the uploaded data, produces responses that are directly relevant to the user's inquiries, facilitating efficient knowledge extraction and decision-making.

Overall, this project represents a significant advancement in leveraging LLMs for practical applications by enhancing their contextual understanding through custom data prompts. By bridging the gap between users and their datasets, the web-based portal empowers users to interact with their data in a more intuitive and productive manner. As a result, it holds the potential to revolutionize various domains ranging from healthcare and education to business intelligence, unlocking new possibilities for leveraging the capabilities of large language models in real-world scenarios.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

INTRODUCTION

This project aims to address a significant limitation of Large Language Models (LLMs) like GPT-3, which is their relatively short context length. LLMs have a restricted capacity to understand and generate responses based on lengthy or complex input data. To overcome this limitation, we propose the development of a web-based portal that seamlessly integrates custom data prompts with LLMs. The core functionality of this portal revolves around enhancing the contextual understanding of LLMs by inserting custom data prompts. These prompts could encompass various forms of data, including text, images, audio, video, and even YouTube videos. By incorporating such diverse data types, the portal ensures a richer and more comprehensive context for the LLM to operate within.

Users of the portal will have the ability to engage in conversational interactions with their data. They can input queries or prompts tailored to their specific dataset, and the LLM, augmented by the integrated custom data, will generate responses that are directly relevant to the input. This approach eliminates the need for users to sift through redundant or extraneous information when seeking answers from their data. For example, a user could upload a collection of medical case studies as custom data. They could then pose questions related to specific patient symptoms, treatment protocols, or outcomes. The LLM, empowered by the contextual information provided by the medical case studies, would be able to generate accurate and informative responses tailored to the user's inquiries.

Overall, this project represents a pioneering effort in leveraging LLMs in conjunction with custom data prompts to enable more effective communication and interaction between users and their datasets. By enhancing the contextual understanding of LLMs, the web-based portal facilitates more precise and relevant responses, ultimately enhancing the utility and usability of large language models in practical applications.

1.1 Problem Statement

Existing methods of leveraging Large Language Models (LLMs) are constrained by the challenge of low context length, limiting their utility in effectively processing and generating coherent responses from extensive datasets or documents. This limitation poses a significant barrier to users seeking to extract relevant insights and answers from large-scale data sources efficiently. Moreover, traditional approaches often result in redundant or irrelevant information in responses, diminishing the user experience and hindering productivity.

Therefore, there is an urgent need to develop an innovative solution that addresses these challenges by creating a web-based portal. This portal will serve as an advanced interface for interacting with LLMs while seamlessly integrating custom data prompts. By doing so, it will empower users to engage in meaningful dialogue with their data across a diverse range of formats, including text, image, audio, and video. Importantly, the project aims to eliminate redundancy in responses, ensuring that users receive concise and targeted information without unnecessary data overload.

In essence, the primary objective of this project is to enable users to communicate with their data in a natural and intuitive manner, leveraging the power of LLMs to provide contextually relevant insights and answers tailored to the user's specific queries and input data.

1.2 Scope

The scope of this project is as follows:

- Build a platform where all the users can upload their personalized data and talk to it without any redundant or web-scraped data.
- This platform also allows the users to overcome the token limitation of LLMs and have various types of data as input (text, pdf, audio, video, image).

The application allows users to:

- One Stop Platform.
- Upload various types of data.
- Talk to your text data.
- Convert text to image.
- Summarize real-time videos.
- Convert audio to text – Transcription and Summarization.

1.3 Motivation

In the landscape of burgeoning chat services, we observed a common trend among companies striving to provide specific replies and adeptly handle edge cases. However, we also noticed a significant barrier: the cost associated with similar services on platforms like Microsoft Azure. As students, we found ourselves constrained by financial limitations, yet determined to innovate. Thus, we set out on a mission to create a solution that not only addressed the need for tailored conversational experiences but also surpassed the token limitations in large language models, all without requiring substantial financial investment. Our project embodies the fusion of necessity and ingenuity – a testament to our unwavering commitment to democratizing access to advanced technology. It’s a journey fueled by resilience, where every obstacle becomes an opportunity for growth. By combining our passion for specific replies with the desire to overcome limitations, we’ve not only created a solution but also a narrative of perseverance and innovation. Our project is more than just a technical endeavor; it's a testament to the power of 5here5ineation, creativit’, and the unwavering belief that even the most audacious dreams are within reach. With this motivation driving us forward, we are poised to make a meaningful impact in the world of conversational AI. Jndckncknckdncksdnckdncknd ckdnck

CHAPTER 2

LITERATURE SURVEY

2.1 Survey of Existing/Similar System

Ref Id	Feature	Gap Analyzed
1	Context Length Management	Efficiently manages context length limitations inherent in Large Language Models (LLMs), ensuring seamless interaction with extensive datasets without truncation issues. This capability bridges the gap in traditional NLP systems 7here passages of text often lead to incomplete or irrelevant responses.
2	Custom Data Integration	Seamlessly integrates custom data prompts into queries, enhancing the relevance and specificity of responses generated by LLMs. By allowing users to provide additional context, VectorShift empowers users to receive more accurate and tailored answers, addressing the gap in conventional LLM interactions where generic responses may lack precision.

3	Response Quality	Prioritizes the delivery of contextually relevant and coherent responses, leveraging advanced algorithms to ensure that generated answers align closely with user queries and input data. This feature addresses the gap in traditional NLP systems where responses may lack relevance or coherence, leading to inefficiencies in data interpretation and decision-making.
4	Multimedia Support	Extends support beyond text-based inputs to include images, audio, and video, enabling users to interact with diverse data types and receive comprehensive responses across various media formats. By accommodating multimedia inputs, VectorShift enhances the versatility and utility of NLP interactions, bridging the gap in conventional systems that may overlook the rich information available in non-textual formats.

5	Redundancy Elimination	Minimizes redundancy in responses, delivering concise and focused information to users. By eliminating unnecessary repetition or duplication, VectorShift enhances the efficiency and effectiveness of data interaction, addressing the gap in traditional NLP systems where users may encounter redundant or irrelevant information.
6	User Experience	Offers a user-friendly interface designed for intuitive interaction, promoting seamless communication with data. From navigating the platform to inputting queries and interpreting responses, VectorShift prioritizes simplicity and clarity, enhancing the overall user experience and facilitating more efficient data interaction. This feature addresses the gap in traditional NLP systems where complex interfaces or cumbersome workflows may hinder user engagement and productivity.

7	Natural Language Interaction	Facilitates natural language interaction between users and data, allowing users to communicate with their data seamlessly without the need for complex query structures or technical expertise. By enabling intuitive communication, VectorShift enhances user accessibility and usability, addressing the gap in traditional NLP systems where language barriers or technical complexities may limit user engagement and adoption.
8	Real-time Processing	Provides real-time processing capabilities, allowing users to receive instant feedback and responses to their queries. This feature enhances user productivity and responsiveness, addressing the gap in traditional NLP systems where latency or delays in processing may impede real-time decision-making and analysis.

9	Contextual Understanding	Employs advanced algorithms to understand and retain context effectively, ensuring that generated responses maintain coherence and relevance to the user's queries and input data. By preserving context, VectorShift enhances the accuracy and reliability of generated answers, addressing the gap in traditional NLP systems where context may be lost or misinterpreted, leading to inaccuracies in responses.
10	Personalization	Allows customization of prompts and preferences to tailor responses to the specific needs and preferences of users. By accommodating individual preferences, VectorShift enhances user satisfaction and engagement, addressing the gap in traditional NLP systems where one-size-fits-all approaches may limit user customization and flexibility.

11	Continuous Improvement	Incorporates feedback mechanisms and iterative refinement processes for ongoing enhancement and optimization. By actively seeking user feedback and iterating on features, VectorShift ensures continuous improvement and adaptation to evolving user needs and industry trends, addressing the gap in traditional NLP systems where stagnation or obsolescence may impede innovation and progress.
12	Analytics and Insights	Generates analytics and insights from user interactions and data inputs, providing users with valuable feedback and actionable intelligence. By leveraging analytics and insights, VectorShift empowers users to derive deeper insights and make informed decisions, addressing the gap in traditional NLP systems where data analysis and interpretation may be limited or overlooked.

Context Length Management:
VectorShift efficiently manages context length limitations inherent in Large Language Models (LLMs). It ensures seamless interaction with extensive datasets without truncation issues, addressing a common challenge in traditional NLP systems.

Custom Data Integration:
Seamlessly integrates custom data prompts into queries, enhancing the relevance and specificity of responses generated by LLMs. This empowers users to receive more accurate and tailored answers, bridging the gap in conventional LLM interactions.

Response Quality:
Prioritizes the delivery of contextually relevant and coherent responses, aligning closely with user queries and input data. By ensuring high-quality responses, VectorShift enhances the efficiency of data interpretation and decision-making processes.

Multimedia Support:
Extends support beyond text-based inputs to include images, audio, and video, enabling comprehensive interactions. By accommodating multimedia inputs, VectorShift enhances the versatility and utility of NLP interactions, addressing gaps in traditional systems.

Redundancy Elimination:
Minimizes redundancy in responses, delivering concise and focused information. By eliminating unnecessary repetition or duplication, VectorShift enhances the effectiveness of data interaction, improving user efficiency and satisfaction.

User Experience:
Offers a user-friendly interface designed for intuitive interaction, promoting seamless communication with data. VectorShift prioritizes simplicity and clarity, enhancing the overall user experience and facilitating more efficient data interaction.

Natural Language Interaction:

Facilitates natural language interaction between users and data, enabling intuitive communication without complex query structures. This enhances user accessibility and usability, bridging language barriers and technical complexities.

Real-time Processing:
Provides real-time processing capabilities, allowing users to receive instant feedback and responses to their queries. This feature enhances user productivity and responsiveness, facilitating real-time decision-making and analysis.

Contextual Understanding:
Employs advanced algorithms to understand and retain context effectively, ensuring coherence and relevance in generated responses. By preserving context, VectorShift enhances the accuracy and reliability of answers.

Personalization:
Allows customization of prompts and preferences to tailor responses to individual needs. By accommodating user preferences, VectorShift enhances user satisfaction and engagement, fostering a personalized user experience.

Integration Flexibility:
Integrates with various platforms and applications, providing seamless access and interoperability. This enhances usability and scalability, supporting organizational growth and collaboration.

Analytics and Insights:
Generates analytics and insights from user interactions and data inputs, providing valuable feedback and actionable intelligence. By leveraging analytics, VectorShift empowers users to derive deeper insights and make informed decisions.

Continuous Improvement:
Incorporates feedback mechanisms and iterative refinement processes for ongoing enhancement. By actively seeking user feedback, VectorShift ensures continuous improvement and adaptation to evolving user needs and industry trends.

Security and Privacy:
Ensures data security and privacy through encryption, access controls, and regulatory compliance. By prioritizing security, VectorShift safeguards user data and instills confidence in data handling and protection.

Scalability:
Scales to accommodate growing data volumes and user demands, ensuring consistent performance and reliability. By offering scalability, VectorShift supports organizational growth and usage expansion.

This non-tabular format provides a detailed overview of VectorShift’s features, highlighting its capabilities and contributions to revolutionizing natural language processing.

Link: https://www.vectorshift.ai/

CHAPTER 3

SYSTEM DESIGN

3.1 PROPOSED SYSTEM ARCHITECTURE

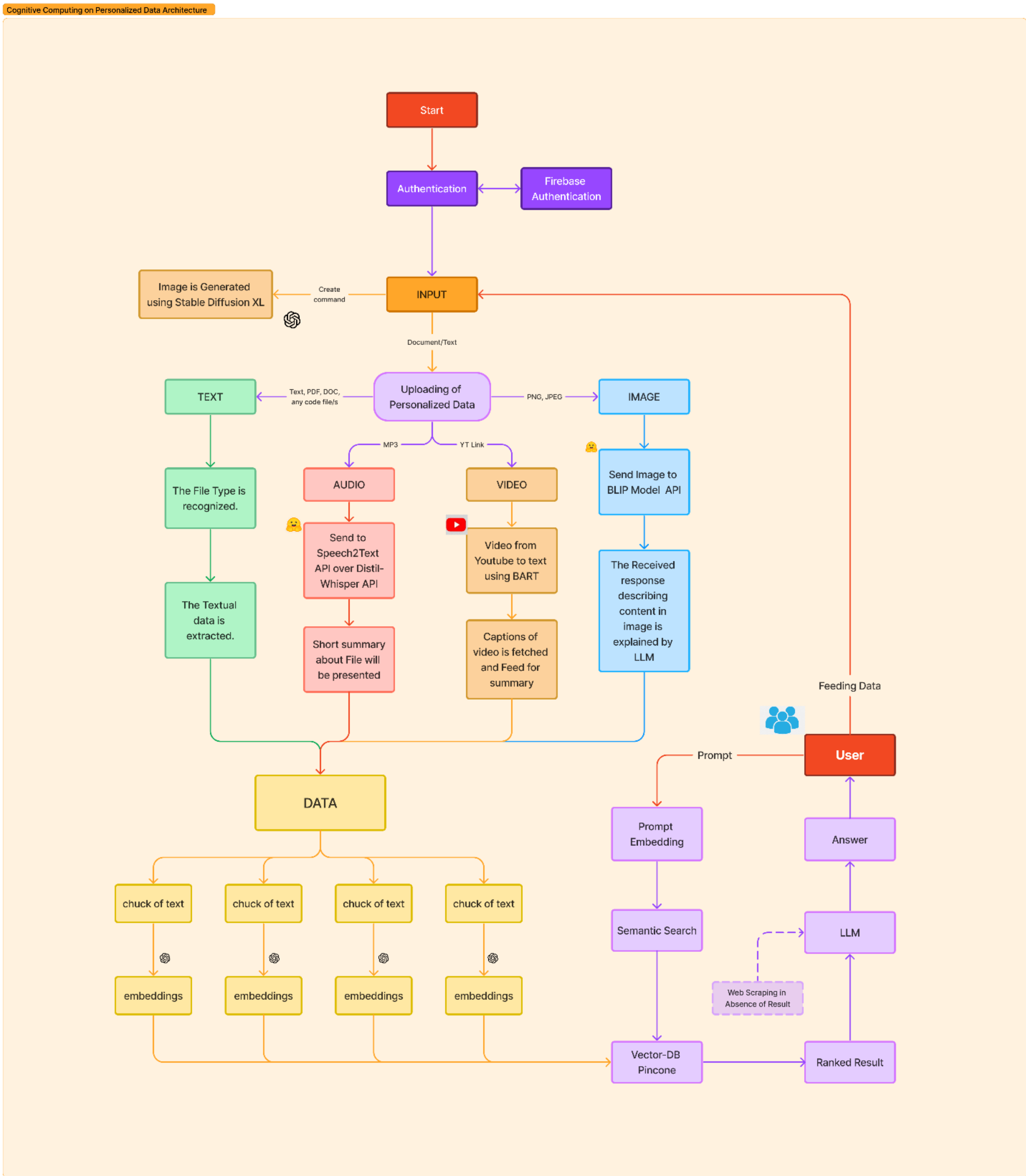


Fig 3.1: Extended System Architecture

3.2 Flowchart Description:

In LangChain, the utilization of a recursive model represents a pivotal innovation in data processing and retrieval. This recursive model acts as the backbone of the system, breaking down complex information into more manageable chunks. By segmenting data in this manner, LangChain enhances efficiency and scalability, enabling seamless storage and retrieval within the Pinecone database. The intricate backend processes, including data segmentation, vector conversion, and database storage, are seamlessly integrated behind a user-friendly interface. This approach shields users from the complexities of the backend operations, ensuring a smooth and intuitive experience.

When users input prompts, they trigger a semantic search process that retrieves relevant data from the database. This retrieved data undergoes further refinement by a large language model, enhancing clarity and relevance before being presented to the user. This iterative process ensures that users receive accurate and meaningful results tailored to their queries.

Beyond semantic search, LangChain offers a range of additional services such as text-to-video, text-to-image, video-to-text, and summarization. These functionalities are made possible through Flask APIs created in the backend, which facilitate seamless integration and interaction with external services.

From the user's perspective, the process is straightforward and user-friendly. Users begin by logging in using the embedded Firebase Google authentication system, ensuring secure access to the platform. Next, users select the type of data they wish to input, providing flexibility and customization options. Users then add prompts specifying the desired actions or analyses to be performed on the input data.

Behind the scenes, the backend processes these prompts using large language models, leveraging advanced natural language processing techniques to generate accurate and relevant outputs. Finally, the refined output is presented to the user on the front-end interface, completing the user interaction loop.

Overall, LangChain's innovative approach to data processing and retrieval, combined with its intuitive user interface and seamless integration of backend services, offers users a powerful tool for harnessing the full potential of their data.

3.3 Methodology:

We have used a recursive model in LangChain. This model breaks information into manageable chunks, which are then converted into vectors and stored seamlessly within the Pinecone database. This complex backend process is hidden behind a user-friendly interface. User prompts trigger a semantic search, retrieving relevant data that is then refined by a large language model for improved clarity before being displayed. Other services provided by the website like, text to video, text to image, video to text and summarization are made possible by the use of Flask APIs created in the backend. The overall process is very simple on the user end. The user has to login using the embedded firebase google authentication system. Next, the user has to choose the type of data they want to use as input. Following this the user has to add prompts regarding what needs to be done using the input data. This prompt is then processed in the backend through Large Language Models and the final output is displayed to the user on the frontend. hjygsdjkbsfekjbskjbsdkfjskjksdjfkjdsnfjksnfkkfjnsfkjsnkfjskfjjsfsj

3.4 Analysis:

3.4.1 Process Model

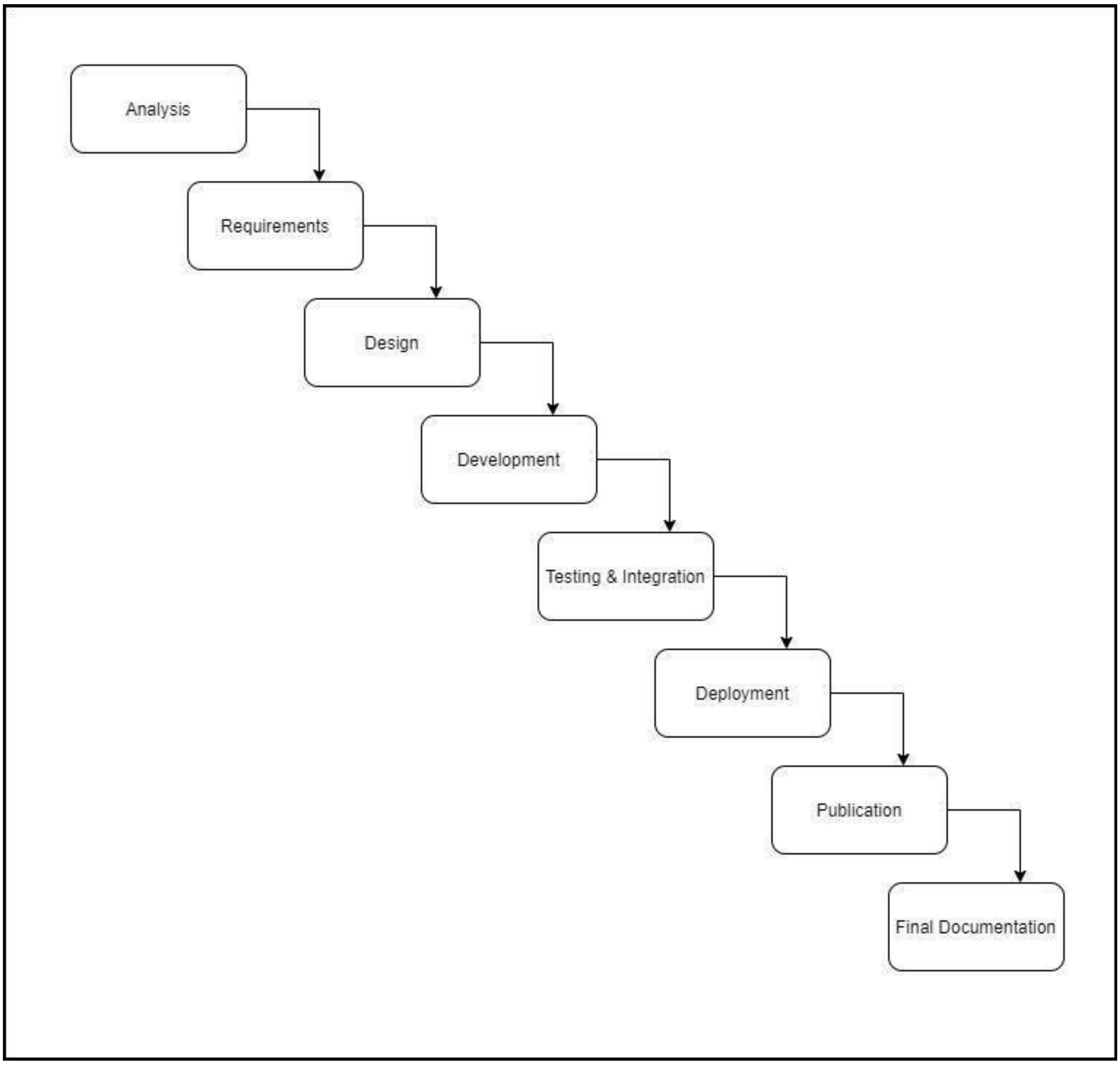


Fig 3.2: Waterfall Model

We intend to use the Waterfall Model in the development of our project. The reasons for using this model in our project are:

- In waterfall, development of one phase starts only when the previous phase is complete. Because of this nature, each phase of the waterfall model is quite precise well defined. Since the phases fall from a higher level to lower level, like a waterfall, it is named as the waterfall model.
- Since the requirements are stable and not changing frequently, the Waterfall Model is bestsuited for the project.
- It enables the monitoring and departmentalization. A timeline can be set with deadlines for each development step, and a product can proceed through the development process modelphases one by one.
- The waterfall model progresses through easily understandable and explainable phases andthus it is easy to use.
- It is easy to manage due to the rigidity of the model – each phase has specific deliverablesand a review process.
- In this model, phases are processed and completed one at a time and they do not overlap.

3.4.2 Feasibility Analysis

1. **Technical feasibility**: Technical feasibility focuses on the technical resources (software and hardware) available to the organization and also helps to determine whether the technical team is capable of converting the ideas into workingsystems. The software required for our project is mainly open-source and readilyavailable (e.g. Langchain, Pinecone, etc.)
2. **Economic feasibility**: This assessment typically involves a cost/ benefits analysis of the project. The project will be developed in minimal cost as most of the software required are open source, the main cost incurred will be of the computer system, the servers to host the application and obviously an OpenAI key.

3. **Legal feasibility**: This assessment investigates whether any aspect of the proposed project conflicts with legal requirements like zoning laws, data protection acts, or social media laws. The project does not involve any legal concerns since all the licences and laws will be respected and included in the project.

4. **Operational feasibility**: This assessment involves undertaking a study to analyze and determine whether—and how well—the organization's needs can be met by completing the project. The main objective of this project is to provide a platform for all the users to surpass the basic token limitation and talk to their data.

3.4.3 Timeline Chart

Event	Start Date	Days to Complete
Project topic search	Jun	10
Project topic selection	Jul	15
Project Planning	Aug	20
Defining Requirements	Sept	25
Project Design	Oct	30
Implementation	Dec	30
Testing	Feb	15
Wrap up	March	2

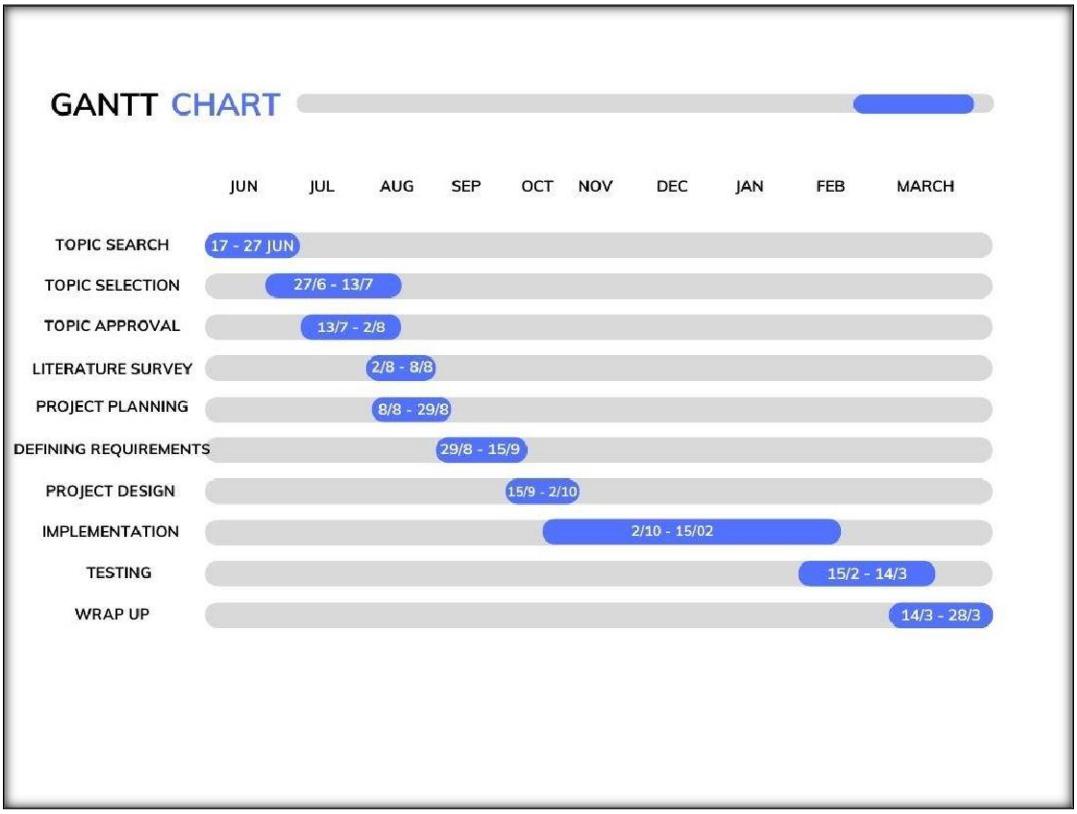


Fig 3.3: Gantt Chart

3.4.4 Design and UML diagram

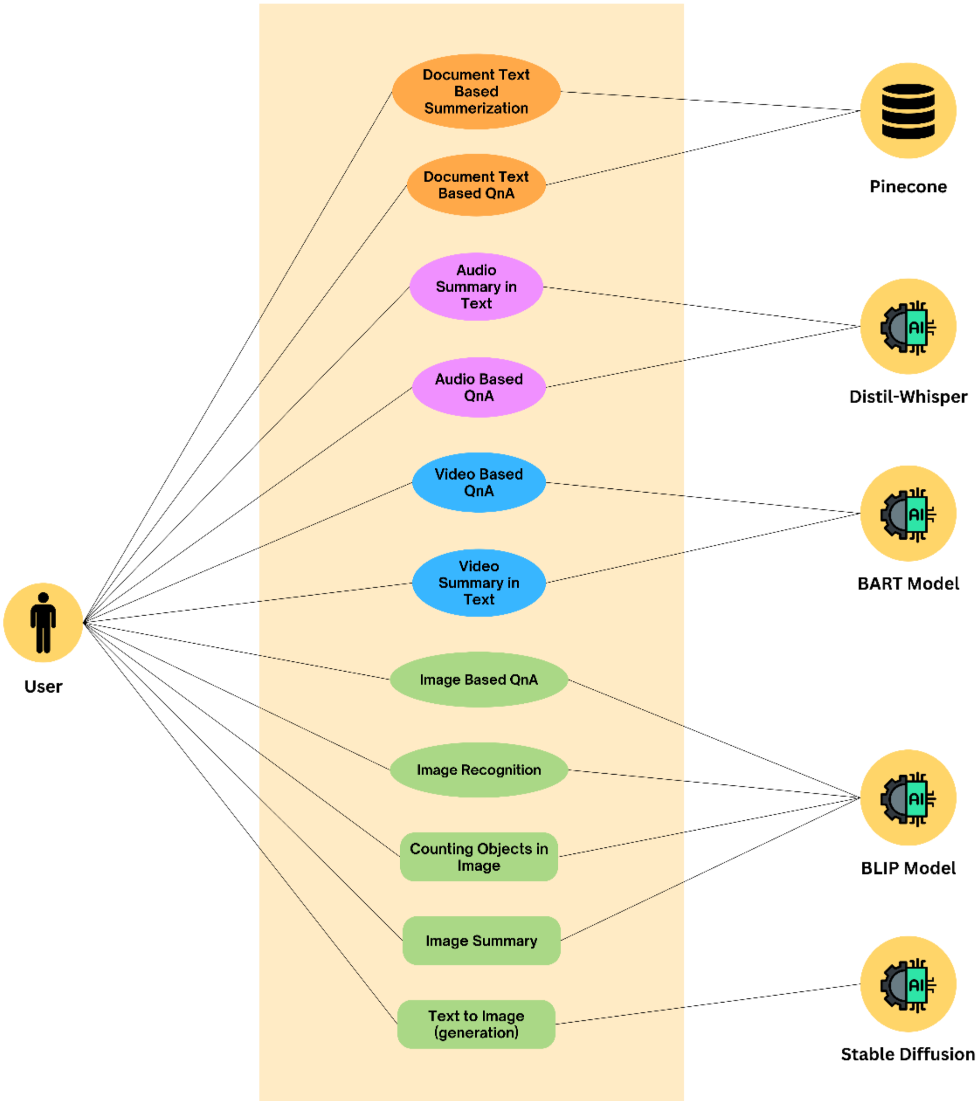


Fig 3.4: UML Diagram

From Fig 3.4, we can see that the User has the access to all the functionalities in the system. Pinecone which is a vector database has access to all the services except the ones which require the OpenFlamingo endpoints. Finally OpenFlamingo endpoints each serve a different function. jbkscdkjsvdbkjbsdkjdsbkjdskjscsjcjkdjdncjnscjcdkjdcjsdckjsc

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 Proposed System

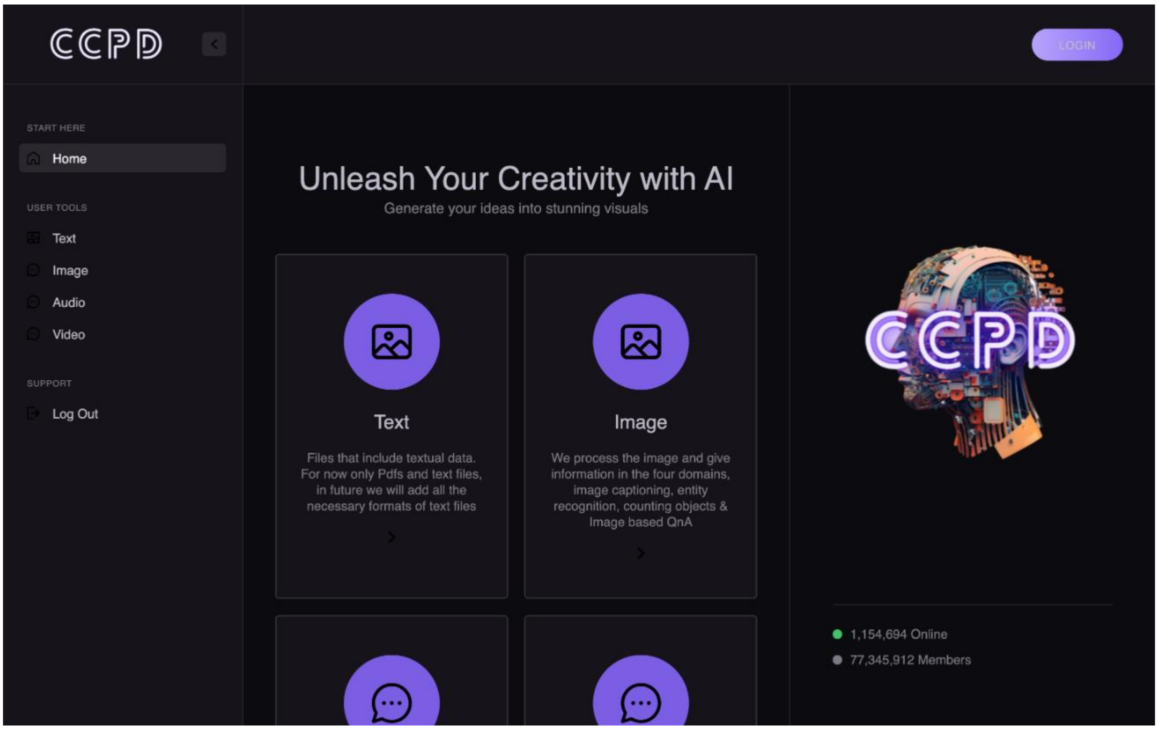
We have used a recursive model in LangChain. This model breaks information into manageable chunks, which are then converted into vectors and stored seamlessly within the Pinecone database. This complex backend process is hidden behind a user-friendly interface. User prompts trigger a semantic search, retrieving relevant data that is then refined by a large language model for improved clarity before being displayed. Other services provided by the website like, text to video, text to image, video to text and summarization are made possible by the use of Flask APIs created in the backend. The overall process is very simple on the user end. The user has to login using the embedded firebase google authentication system. Next, the user has to choose the type of data they want to use as input. Following this the user has to add prompts regarding what needs to be done using the input data. This prompt is then processed in the backend through Large Language Models and the final output is displayed to the user on the frontend.

Proposed Features:

- One Stop Platform.
- Upload various types of data.
- Talk to your text data.
- Convert text to image.
- Summarize real-time videos.
- Convert audio to text – Transcription and Summarization.

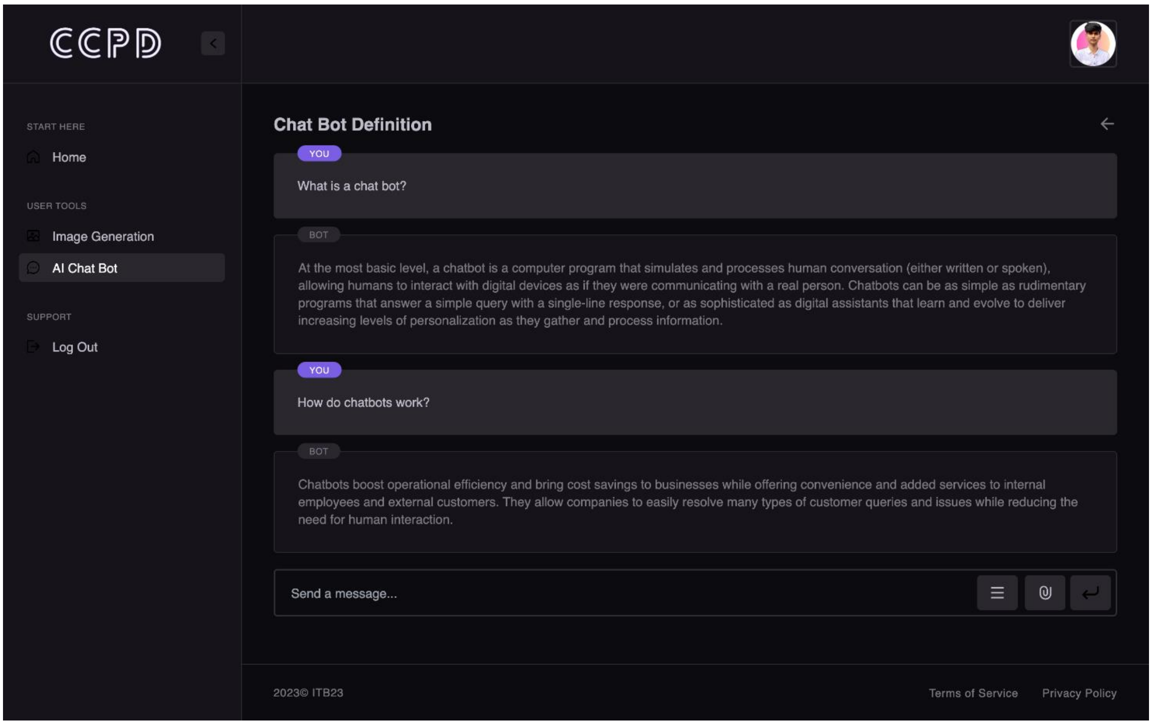
4.2 Features

A. Welcome Screen:



4.1 Welcome Screen

B. Main Chat Portal

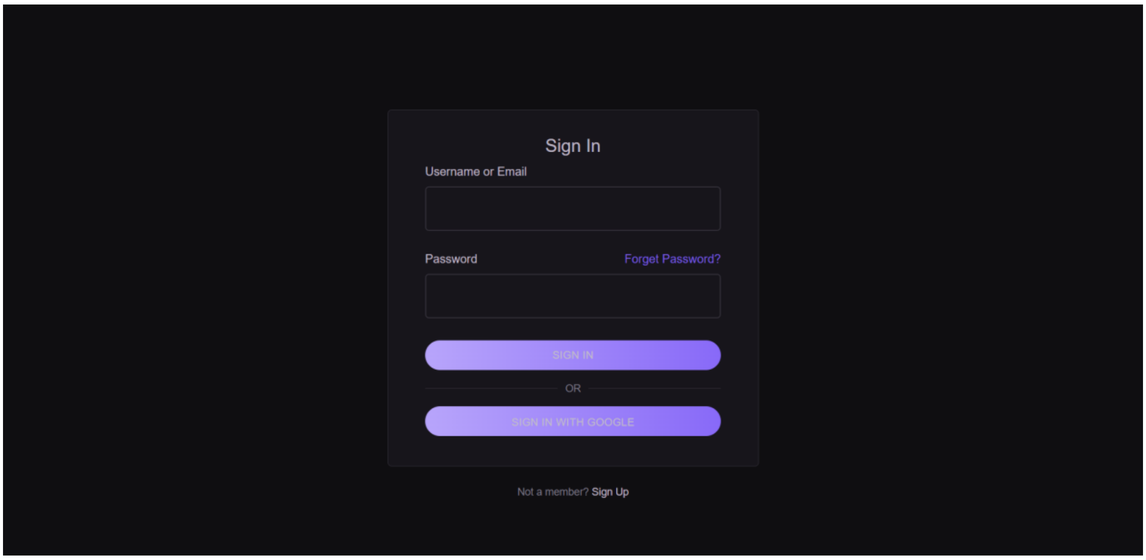


4.2 Main chat portal

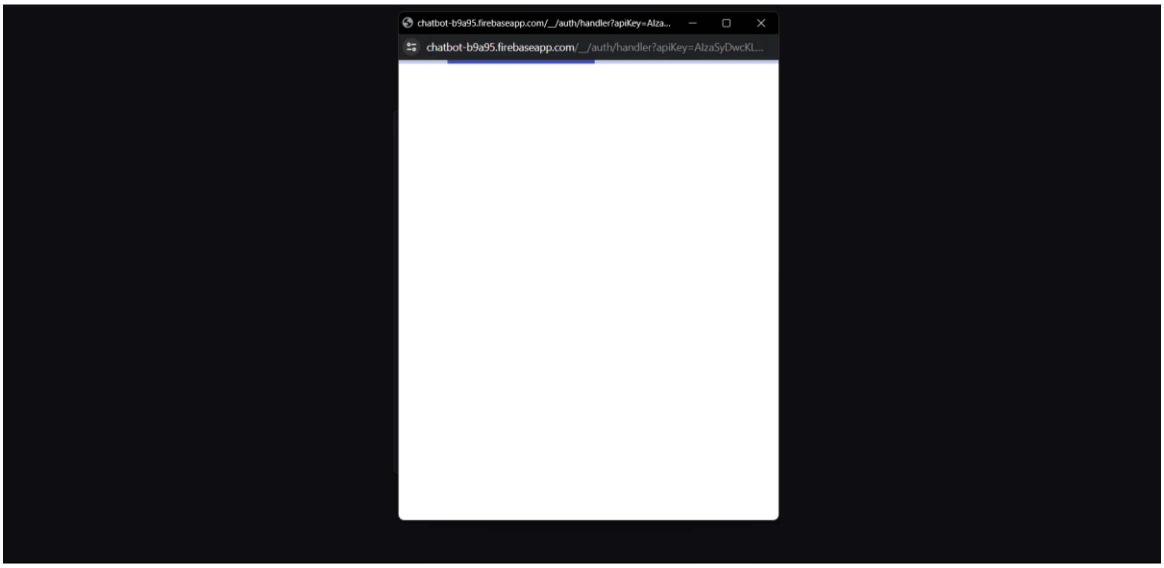
C. Login System:

a) User Registration: The application will allow users to register themselves and create profiles.

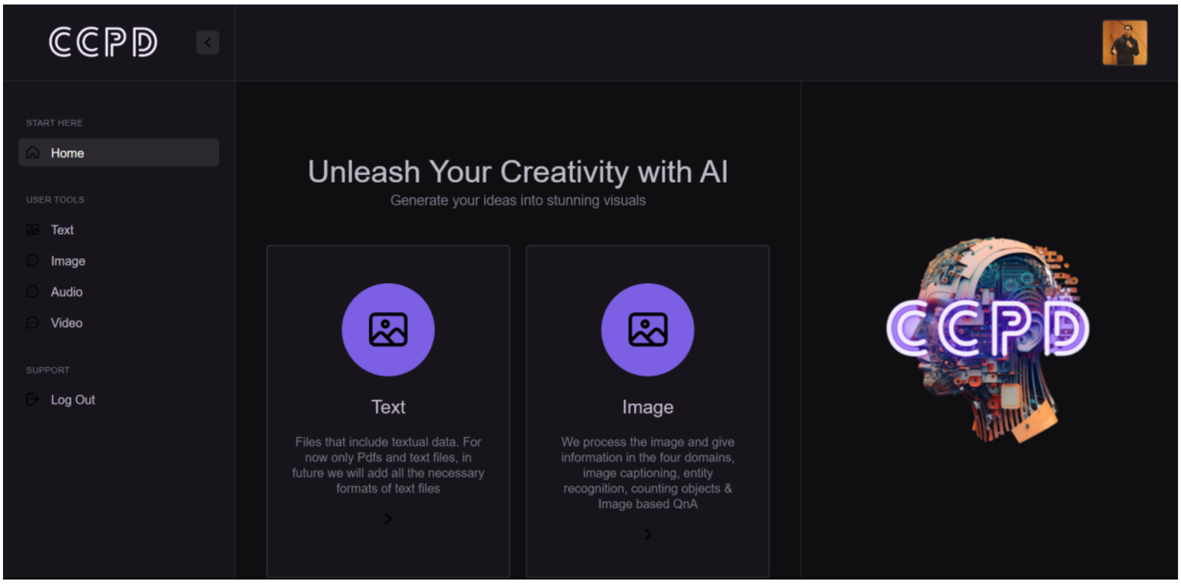
b) Firebase: The application also allows users to directly authenticate using Google accounts.



4.3 Login Screen



4.4 User Login using Firebase



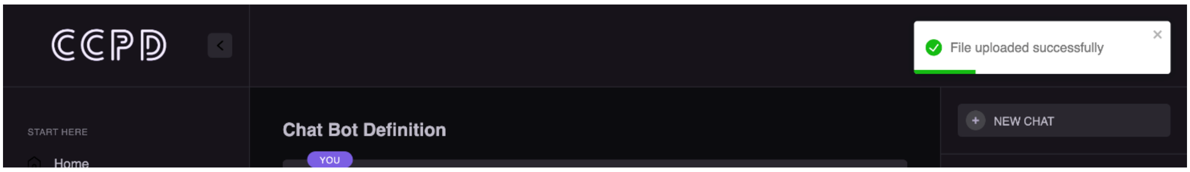
4.5 User Login complete

Text Input:

4.2.1. Various use cases of text input:

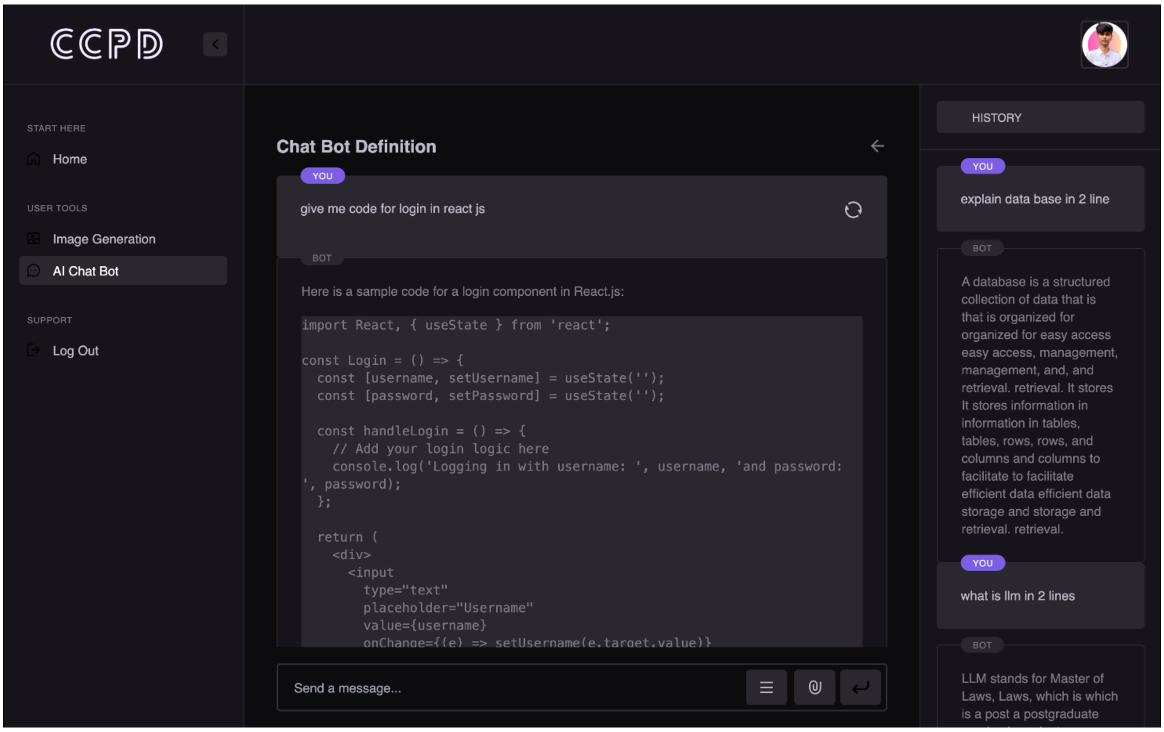
This is the USP of the project.

In this input, the user needs to add text data as input to which they can talk to. Essentially, this is where we surpass the token limitation as well as character limitation.



4.6 File upload

Whenever user uploads a document, they will be able to see a pop up in the top-right corner of the screen.



4.7 Code Output

In this case, we have uploaded our old project. Then, a prompt is given to the chat window for which an appropriate response is shared. Since we have attached the famous OpenAI LLM, we can ask the chat window to correct the code or enhance the code.

This functionality is not limited to code, we can even add a pdf and ask questions for which specific responses from the pdf are answered without any redundant data.

In addition, users can use this system as general LLM as well. But with our project, the user can surpass any amount of token limitation.

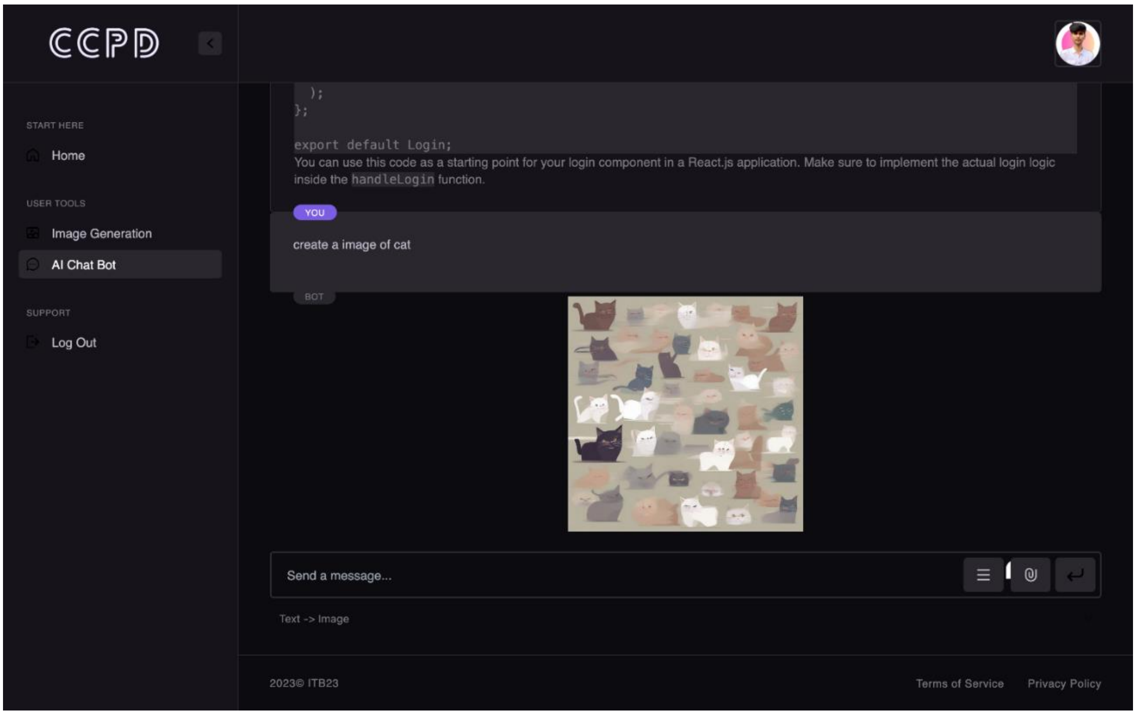
We have also added the history functionality which stores the history of the previous chats. This history is account and session based.

4.2.2. Text to Image:

Text-to-image conversion with Hugging Face entails utilizing pre-trained models, such as CLIP (Contrastive Language-Image Pre-training), which have been trained on large-scale datasets to understand the relationship between text and images. These models can generate images based on textual descriptions by learning the correlations between different modalities.

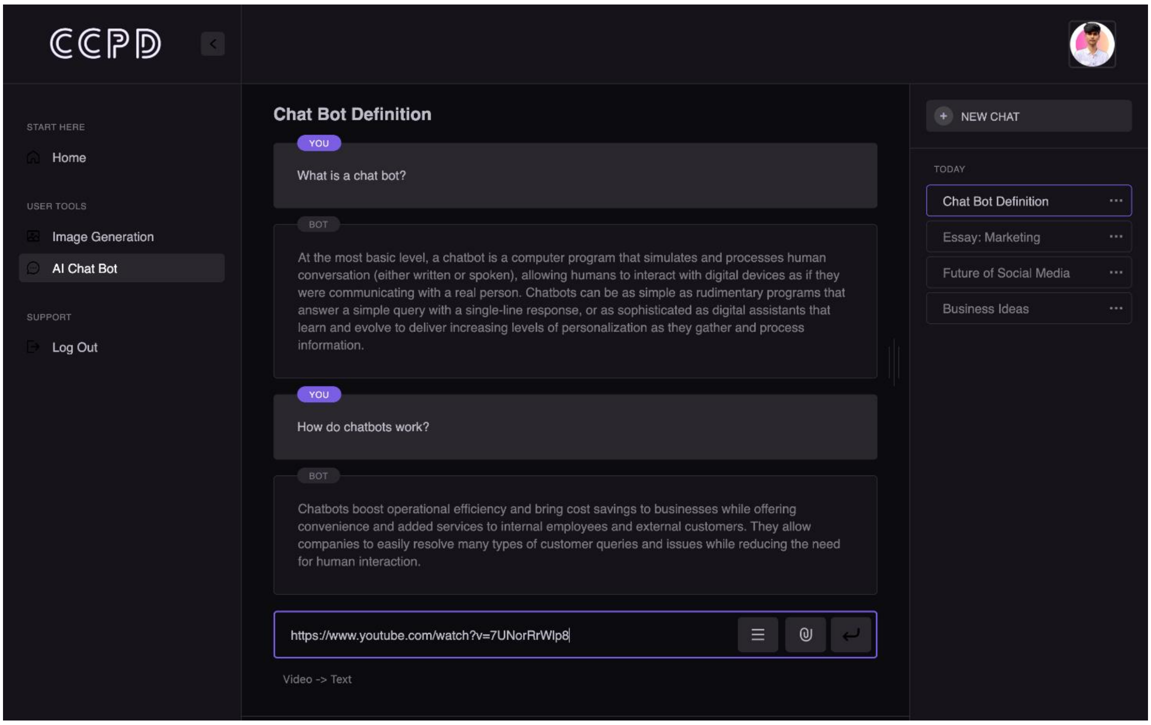
When a user provides a textual input, the model processes the text and uses its understanding of the semantics and context to generate an image that best corresponds to the description. This process involves encoding the text into a vector representation and then decoding it to produce an image.

The advantage of using Hugging Face for text-to-image conversion lies in the availability of state-of-the-art pre-trained models and easy-to-use interfaces, allowing developers to integrate this functionality into their applications with minimal effort.

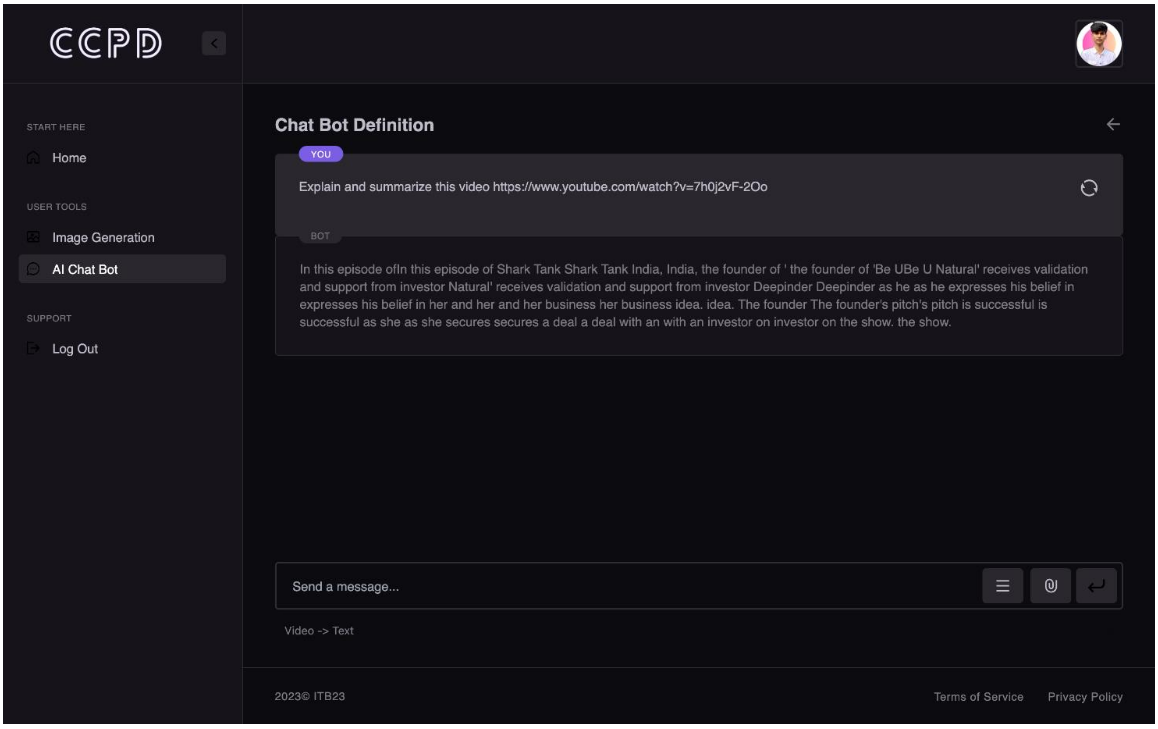


4.8. Text to Image

4.2.3. Video to Text / Text to Video:

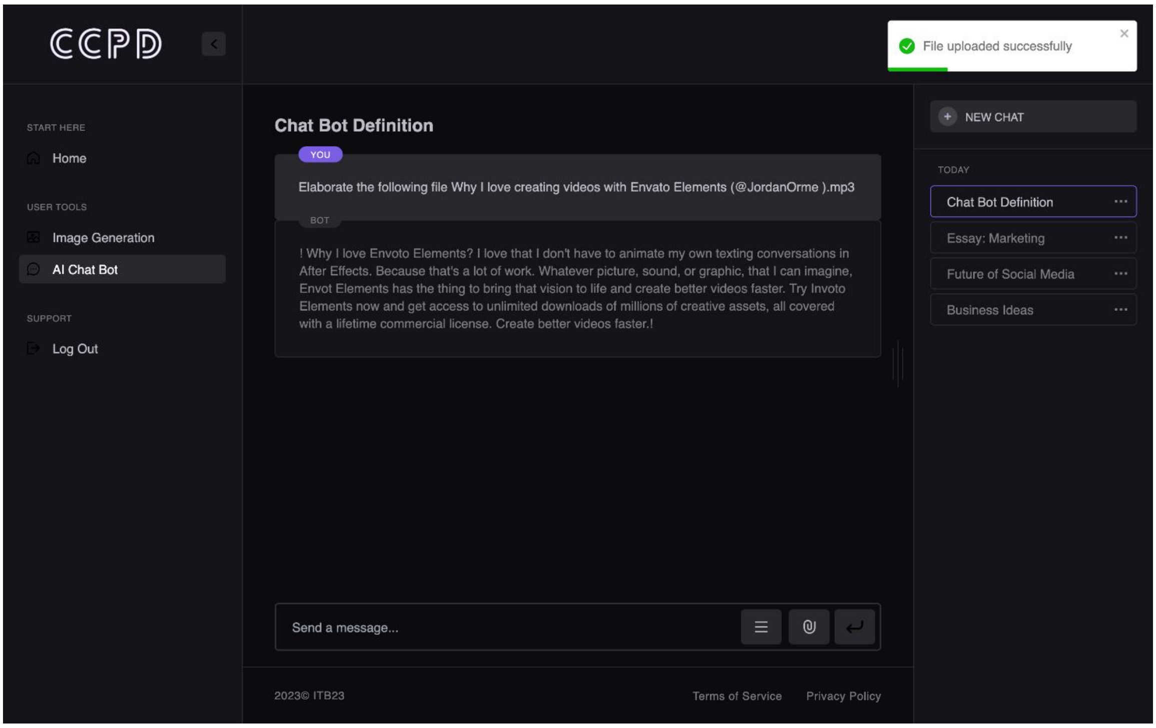


4.9 Add YouTube video link



4.10 YouTube video summarized

4.2.4. Audio to Text:

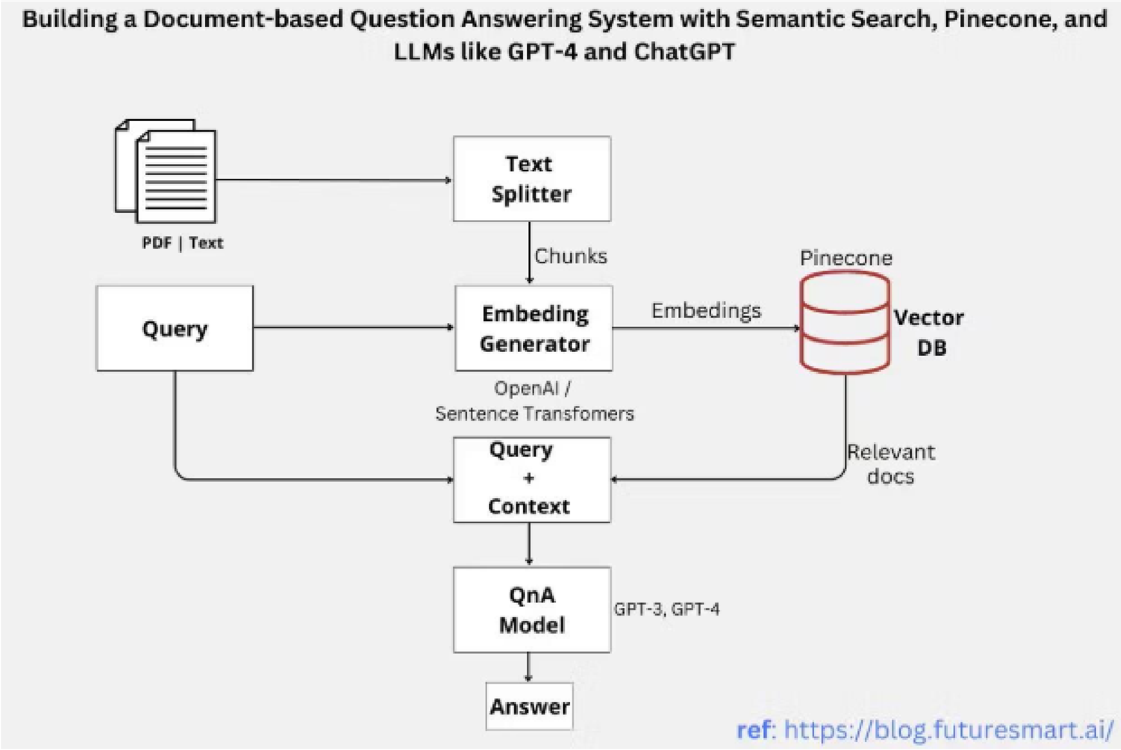


4.11 Audio uploaded and transcribed

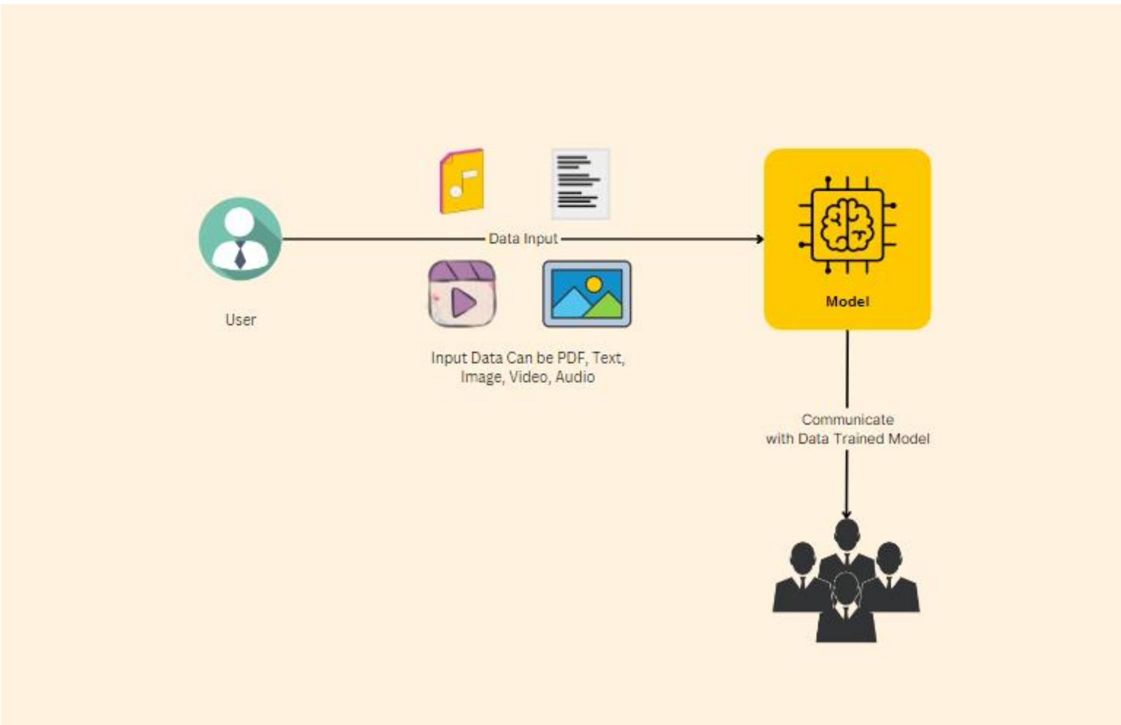
CHAPTER 5

DESIGN

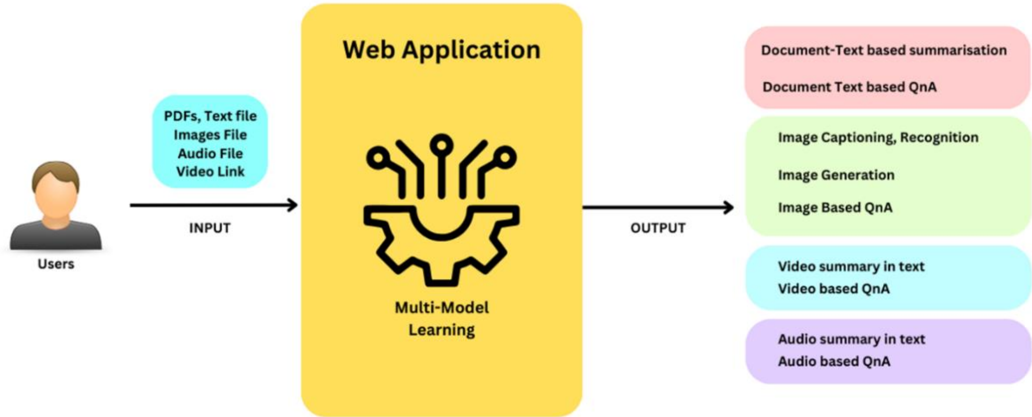
5.1 Data Flow



5.1 DFD 0

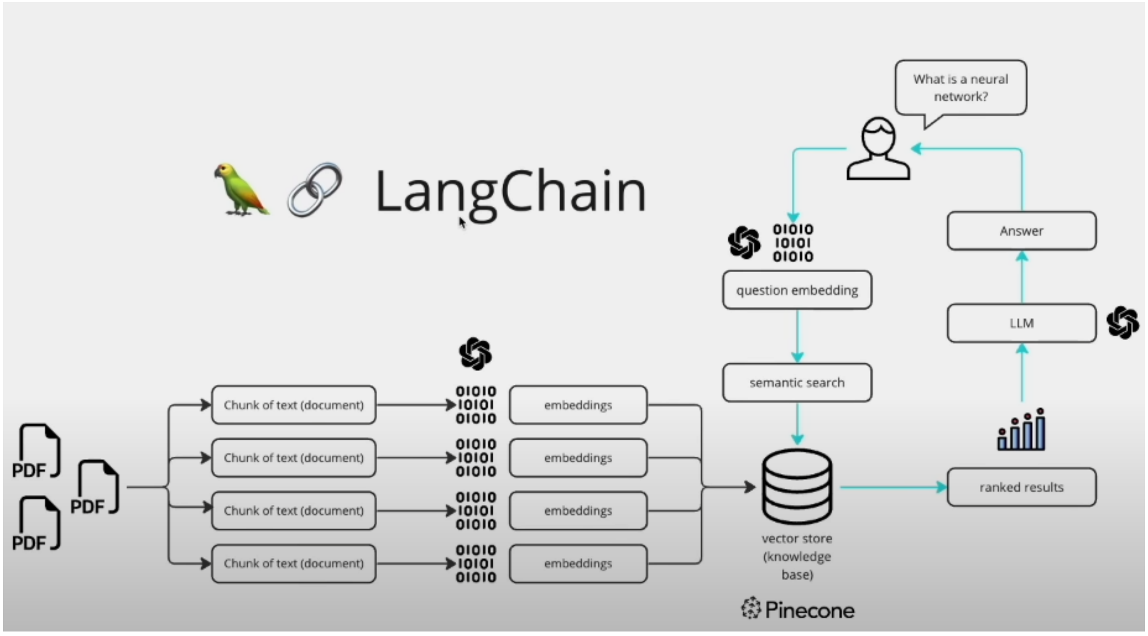


5.2 DFD 1


Web Application
PDFs, Text file
Images File
Audio File
Video Link
Users
INPUT
Multi-Model
Learning
OUTPUT
Document-Text based summarisation
Document Text based QnA
Image Captioning, Recognition
Image Generation
Image Based QnA
Video summary in text
Video based QnA
Audio summary in text
Audio based QnA

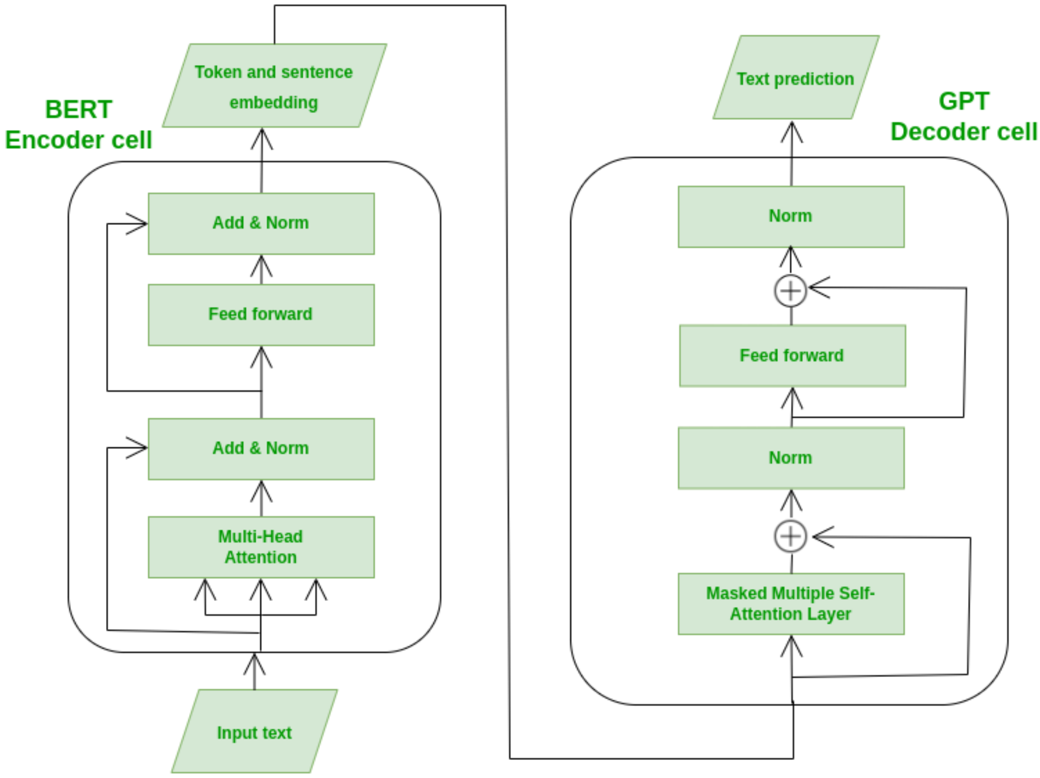
5.3 DFD 2

5.2 Framework Architecture



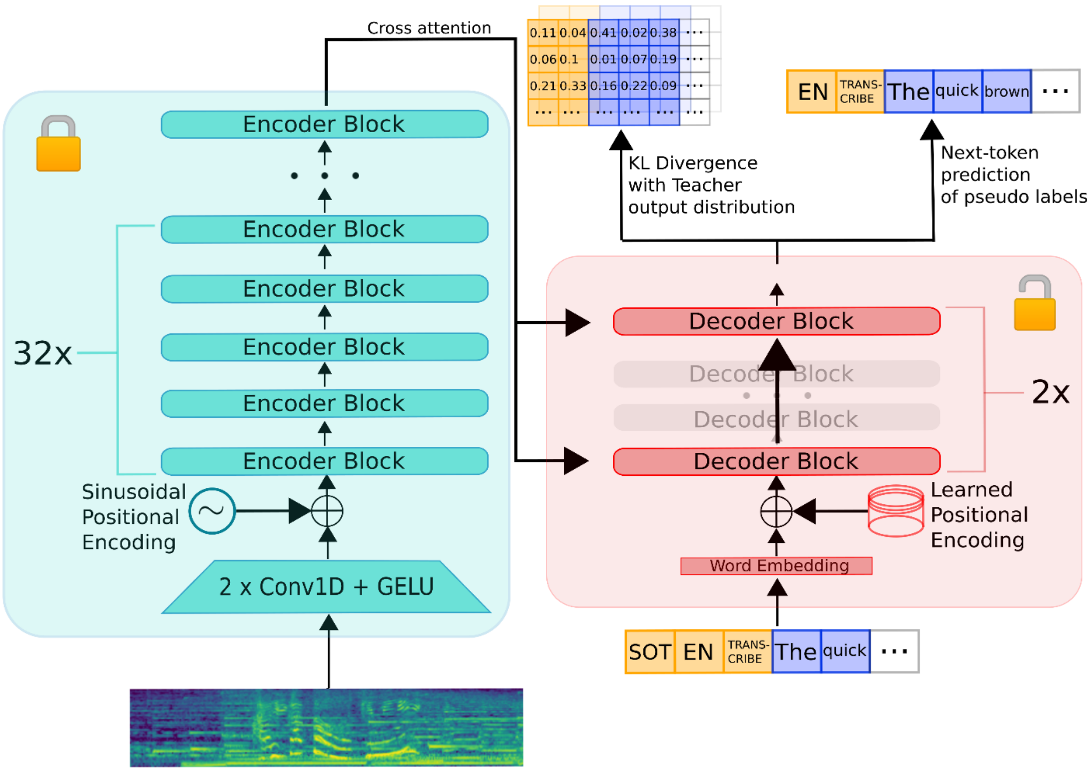
5.4 Langchain Architecture

LangChain's architecture revolves around leveraging cutting-edge AI algorithms to offer personalized language learning experiences. It comprises modular components and off-the-shelf chains for easy customization and higher-level tasks. The platform prioritizes community engagement through language exchange and fosters continuous innovation to meet users' evolving needs. Potential collaborations with OpenAI could further enhance its capabilities, positioning LangChain as a leading contender in shaping the future of language education.



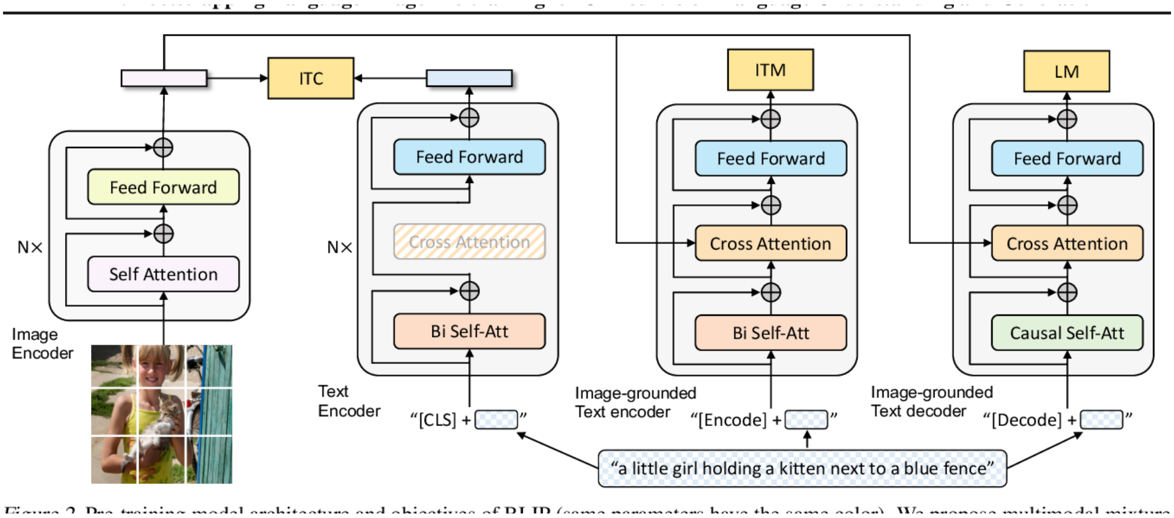
5.5.1 BART Architecture

BART is a denoising autoencoder with a hybrid architecture combining aspects of BERT and GPT models. It features a bi-directional encoder like BERT and an autoregressive decoder like GPT. The encoder generates embeddings for input tokens and a sentence-level embedding, while the decoder learns to generate clean text from corrupted input. BART's architecture includes multi-head attention, addition and normalization, and feed-forward layers. It's designed for tasks like Natural Language Generation and Translation, utilizing masked language modeling for training.



5.5.2 Distil-Whisper Architecture

Distil-Whisper is a compact variant of the Whisper model designed for English speech recognition. It utilizes an encoder-decoder architecture, inherited from Whisper, where the encoder maps speech inputs to hidden-state vectors, and the decoder predicts text tokens autoregressively. During training, Distil-Whisper undergoes knowledge distillation via large-scale pseudo-labelling, using predictions from the Whisper model. It achieves comparable performance to Whisper but with smaller size and faster inference, particularly optimized for both short and long-form transcriptions.



5.5.3 BLIP Architecture

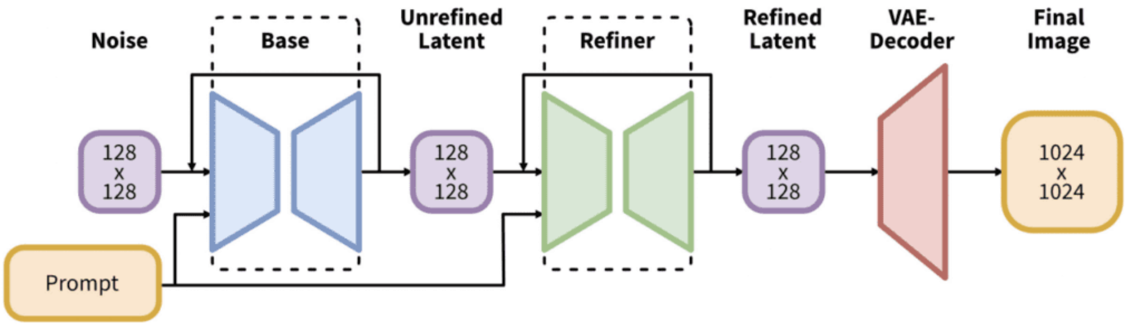
BLIP achieves state-of-the-art performance on a wide range of vision-language tasks, including image-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialog. BLIP integrates vision and language tasks through three key components:

Text Encoder: Encodes text using a Masked Language Model, summarizing with a [CLS] token.

Image-grounded Text Encoder: Incorporates vision-language interactions with cross-attention layers, distinguishing positive and negative pairs.

Image-grounded Text Decoder: Generates text from images using causal self-attention, trained with Language Modeling loss.

These components, trained with multimodal objectives, achieve state-of-the-art performance in tasks like image captioning.



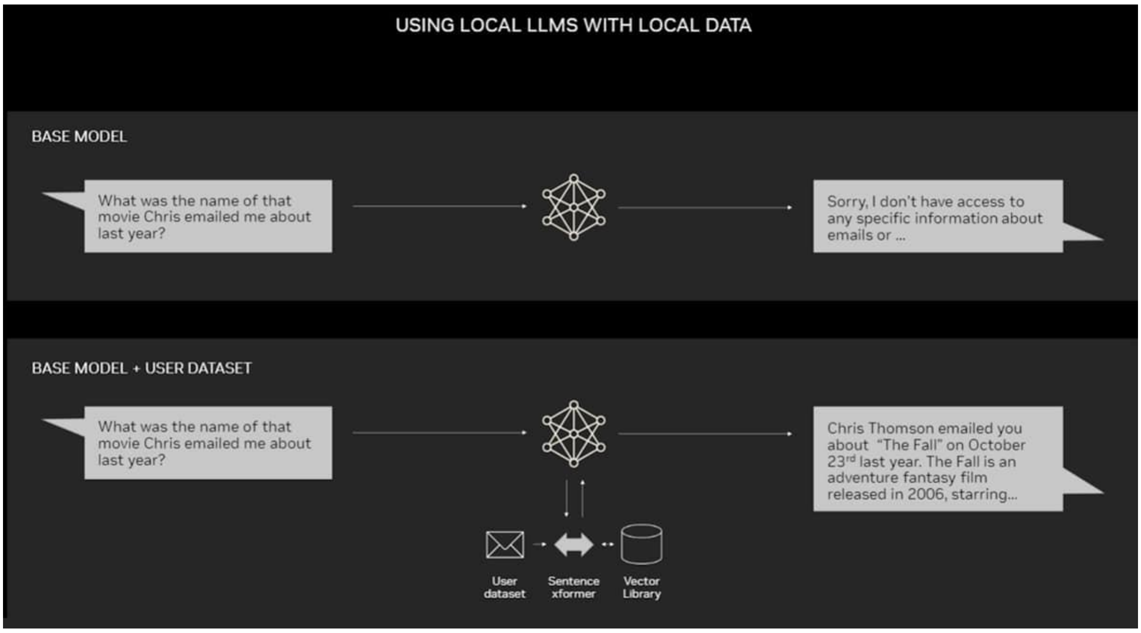
5.5.4 Stable Diffusion Architecture

Stable Diffusion XL is a powerful text-to-image model designed to create incredibly detailed images. It achieves this through two key features:

1. **Supercharged image processing:** At its core, SDXL uses a UNet architecture, which is like a specialized image analysis tool. Unlike previous versions, SDXL's UNet is three times larger, boasting 3.5 billion parameters. This allows it to analyse text prompts and generate images with much finer details compared to earlier models.

2. **Tag-team approach:** SDXL takes a unique approach to image generation. Imagine a two-step process: first, a "base model" analyses the text prompt and creates a draft image. Then, an optional "refiner model" takes over, polishing the draft and adding even more detail. This teamwork between models ensures high-quality results.

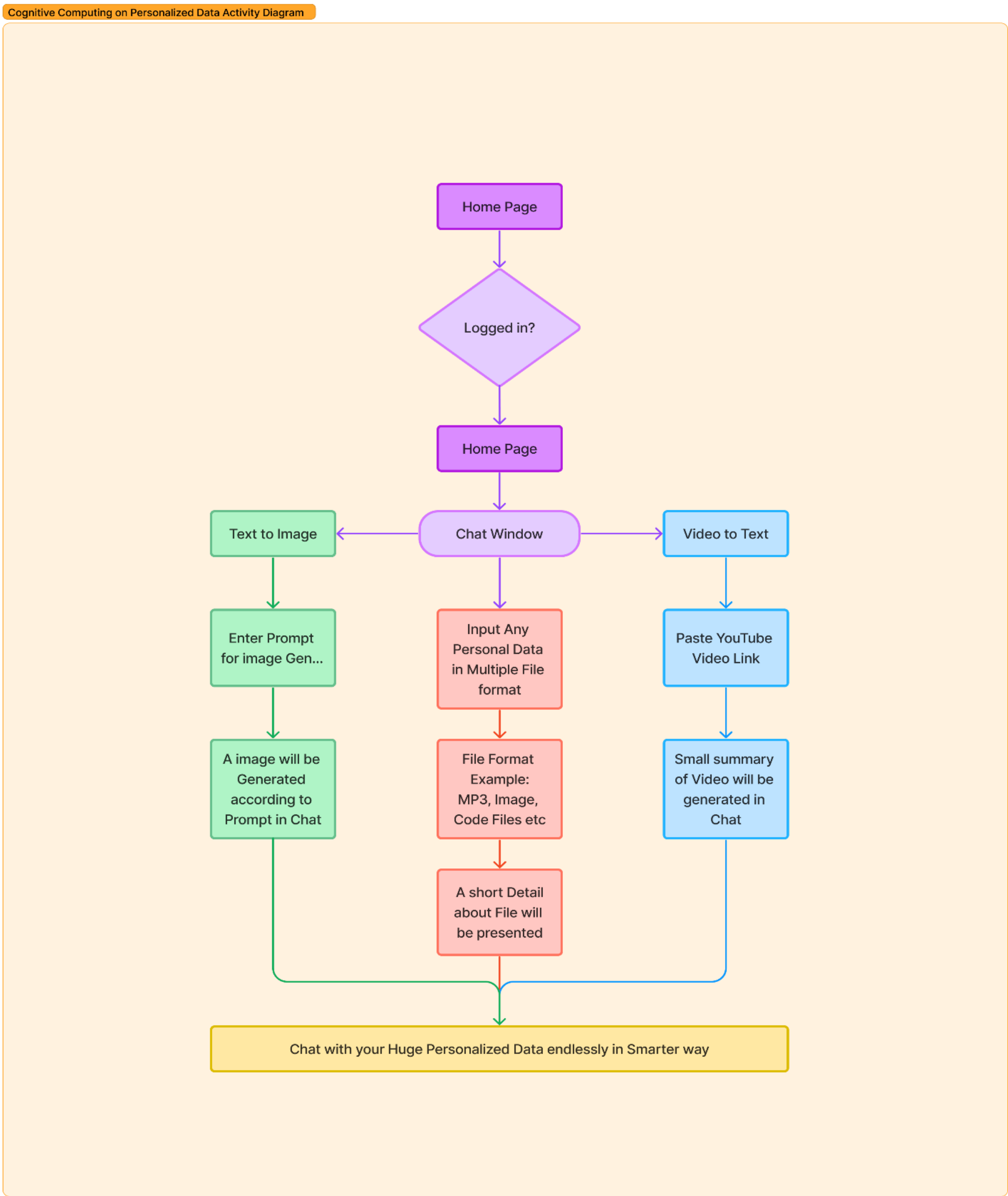
In essence, Stable Diffusion XL is like a highly skilled artist with a powerful toolbox and a meticulous assistant, working together to create stunningly detailed images based on your descriptions.



5.6 Retrieval Augmented Generation

Retrieval augmented generation, or RAG, is an architectural approach that can improve the efficacy of large language model (LLM) applications by leveraging custom data. This is done by retrieving data/documents relevant to a question or task and providing them as context for the LLM. RAG has shown success in supporting chatbots and Q&A systems that need to maintain up-to-date information or access domain-specific knowledge.

5.3 Activity Diagram



5.7 Activity Diagram

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

While the use of LLMs is becoming more and more common these days, there is little to no effort made to overcome limitations of tokens. The only way in the real world is to buy LLM APIs with a higher number of tokens. We encountered this problem during one of our other projects and decided to make a free of cost system that resolves this issue along with making a one of kind system that allows users to talk to their own data. This service is extremely beneficial in the current world. Users can use this system to talk to their data which can be health reports, account statements and company-oriented documentation. Companies can use this system to make a custom chatbot for their web portals. The technologies used to make this project are extremely new and will have no problem of depreciation whatsoever in the future.

6.2. Future Scope

- Develop features for user personalization and customization, allowing users to tailor the system's behavior, preferences, and recommendations based on their individual needs and preferences.
- Create node packages for the entire project for easy installation on different machines.
- Implement robust data privacy and security measures, including end-to-end encryption, access controls, and compliance with data protection regulations, to safeguard user data and ensure trust and confidentiality.
- Integration with various new age technologies like AR, VR, etc.

REFERENCES

REFERENCES

[1] Abid Hassan, M. D. Iftekhar Ali, Rifat Ahammed, Sami Bourouis Mohammad Monirujjaman Khan, “Development of NLP-Integrated Intelligent Web System” December 2021.

[2] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Korn-blith, S., Koh, P. W., Ilharco, G., Wortsman, M., & Schmidt, L. (2023). OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. ArXiv. /abs/2308.01390.

[3]https://blog.futuresmart.ai/building-a-document-based-question-answering-system-withlangchain-pinecone-and-llms-like-gpt-4-and-chatgpt

[4] MED-FLAMINGO: A MULTIMODAL MEDICAL FEWSHOT LEARNER Michael Moor∗1 Qian Huang∗1 Shirley Wu1 Michihiro Yasunaga1 Cyril Zakka, Yash Dalmia1 Eduardo Pontes Reis, Pranav Rajpurkar, Jure Leskovec

[5] MultiModal-GPT: A Vision and Language Model for Dialogue with HumansTao Gong Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, Kai Chen

[6] UnIVAL: Unified Model for Image, Video, Audio and Language Tasks, Mustafa Shukor, Corentin Dancette, Alexandre Rame, Matthieu Cord.

[7] OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision Language Models, Anas Awadalla Irena Gao Josh Gardner Jack Hessel Yusuf Hanafy Wanrong Zhu Kalyani Marathe Yonatan Bitton Samir Gadre Shiori Sagawa Jenia Jitsev Simon Kornblith Pang Wei Koh Gabriel Ilharco Mitchell Wortsman Ludwig Schmidt.

[8] Yingxu Wang, The University of Calgary, “On Cognitive Computing”

[9] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, Ajmal Mian “A Comprehensive Overview of Large Language Models” (Preprint)

[10] OpenAI Tokenizer Playground

[11] Russell Kohn, “Mastering Token Limits and Memory in ChatGPT and other Large Language Models”

[12] GPT-4 Technical Report

[13] Patterns for Building LLM-based Systems & Products

[14] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, Chuang Gan "PrincipleDriven Self-Alignment of Language Models from Scratch with Minimal Human Supervision"

[15] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, Tomas Pfister "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes"

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever "Improving Language Understanding by Generative Pre-Training"

PAPER PUBLISHED

CERTIFICATIONS

IEEE-WIE CRCE

Fr. Conceicao Rodrigues College of Engineering

PRAKALP 2.0

Certificate of Achievement

PRESENTED TO

AYUSH TRIPATHI

For securing SECOND PLACE in the PRAKALP 2.0 Project Competition organised by IEEE-WIE on 16th April 2024.

Dr. Surendra Singh Rathod	**Prof. Swapnali Makdey**	**Prof. Sushma Nagdeote**
(Principal, Fr. CRCE)	(Faculty Incharge, IEEE CRCE)	(Faculty Incharge, WIE CRCE)

IEEE-WIE CRCE

Fr. Conceicao Rodrigues College of Engineering

PRAKALP 2.0

Certificate of Achievement

PRESENTED TO

HASNAIN SAYYED

For securing SECOND PLACE in the PRAKALP 2.0 Project Competition organised by IEEE-WIE on 16th April 2024.

Dr. Surendra Singh Rathod	**Prof. Swapnali Makdey**	**Prof. Sushma Nagdeote**
(Principal, Fr. CRCE)	(Faculty Incharge, IEEE CRCE)	(Faculty Incharge, WIE CRCE)

IEEE-WIE CRCE

Fr. Conceicao Rodrigues College of Engineering

PRAKALP 2.0

Certificate of Achievement

PRESENTED TO

BHARGAV MAHAJAN

For securing SECOND PLACE in the PRAKALP 2.0 Project Competition organised by IEEE-WIE on 16th April 2024.

Dr. Surendra Singh Rathod	Prof. Swapnali Makdey	Prof. Sushma Nagdeote
(Principal, Fr. CRCE)	(Faculty Incharge, IEEE CRCE)	(Faculty Incharge, WIE CRCE)

IEEE-WIE CRCE

Fr. Conceicao Rodrigues College of Engineering

PRAKALP 2.0

Certificate of Achievement

PRESENTED TO

DARSHAN MAHANKAR

For securing SECOND PLACE in the PRAKALP 2.0 Project Competition organised by IEEE-WIE on 16th April 2024.

Dr. Surendra Singh Rathod	Prof. Swapnali Makdey	Prof. Sushma Nagdeote
(Principal, Fr. CRCE)	(Faculty Incharge, IEEE CRCE)	(Faculty Incharge, WIE CRCE)

CERTIFICATE

OF

PARTICIPATION

This is to certify that

Ayush Tripathi

of Information Technology has presented a project titled *"Cognitive Computing on Personalized Data"* in the Tantravihar- 2024, conducted on 23rd April 2024 at Vidyalankar Institute of Technology, Mumbai.

Prof. Geeta Narayanan	***Dr. Saurabh Mehta***	***Dr. Sangeeta Joshi***
Convener	***Organizing Secretary***	***Principal (Acting)***

CERTIFICATE

OF

PARTICIPATION

This is to certify that

Hasnain Sayyed

of Information Technology has presented a project titled *"Cognitive Computing on Personalized Data"* in the Tantravihar - 2024, conducted on 23rd April 2024 at Vidyalankar Institute of Technology, Mumbai.

Prof. Geeta Narayanan	***Dr. Saurabh Mehta***	***Dr. Sangeeta Joshi***
Convener	***Organizing Secretary***	***Principal (Acting)***

CERTIFICATE

OF

PARTICIPATION

This is to certify that

Bhargav Mahajan

of Information Technology has presented a project titled *"Cognitive Computing on Personalized Data"* in the Tantravihar - 2024, conducted on 23rd April 2024 at Vidyalankar Institute of Technology, Mumbai.

Prof. Geeta Narayanan
Convener

Dr. Saurabh Mehta
Organizing Secretary

Dr. Sangeeta Joshi
Principal (Acting)

CERTIFICATE

OF

PARTICIPATION

This is to certify that

Darshan Mahankar

of Information Technology has presented a project titled *"Cognitive Computing on Personalized Data"* in the Tantravihar- 2024, conducted on 23rd April 2024 at Vidyalankar Institute of Technology, Mumbai.

Prof. Geeta Narayanan
Convener

Dr. Saurabh Mehta
Organizing Secretary

Dr. Sangeeta Joshi
Principal (Acting)

Github Link:

https://github.com/ayush-t02/Cognitive_Computing_on_Personalized_Data

Plagiarism Report

tripathi_28042024

ORIGINALITY REPORT

14%	13%	2%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	8%
2	www.softwaretestinghelp.com Internet Source	1%
3	www.jetir.org Internet Source	1%
4	www.simplilearn.com Internet Source	1%
5	Submitted to Southern New Hampshire University - Continuing Education Student Paper	1%
6	byte-project.eu Internet Source	<1%
7	zenodo.org Internet Source	<1%
8	nerdsmagazine.com Internet Source	<1%
9	www.grin.com Internet Source	<1%

10	Submitted to Kaplan International Colleges Student Paper		<1 %
11	ebin.pub Internet Source		<1 %
12	Khoi Dinh Tran, Dat Viet Bui, Ngoc Hoang Luong. "Evolving Prompts for Synthetic Image Generation with Genetic Algorithm", 2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2023 Publication		<1 %

Exclude quotes On

Exclude bibliography On

Exclude matches < 14 words