Customer Support System SFBU

A Master's Project report submitted to
School of Engineering
In Fulfillment for the Degree of
Masters of Science in Computer Science
CS589 Generative AI

A genAI powered chatbot

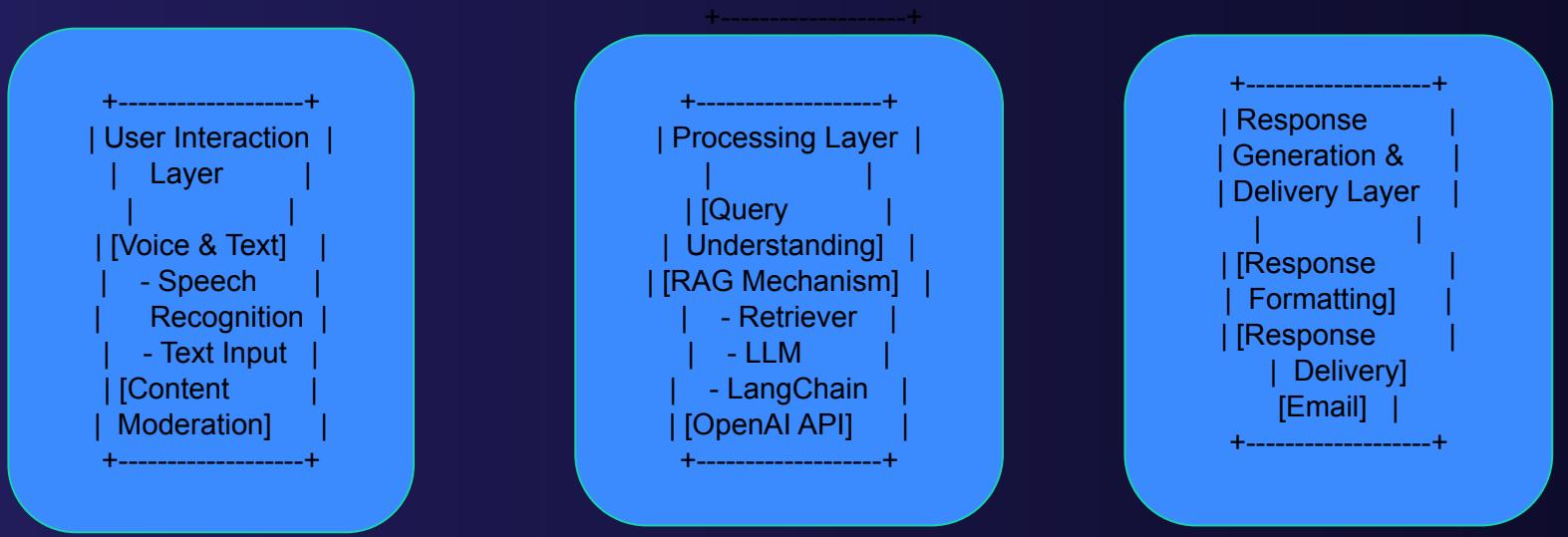
Submitted By-
Professor Adam Weng
Chenxin Cao - 19940
Minh Khoi Duong - 19610
Prachi Sethi - 19963
Prepared under the supervision and guidance of
Professor Henry Chang

INTRODUCTION

- Why GenAI?
- Innovative tool designed to improve communication and accessibility within the school community.
- Cutting-edge AI technology from OpenAI.
- Speech Recognition
- Provides multilingual support

Presented By-
Prachi Sethi

ARCHITECTURE



Presented By-
Prachi Sethi

TECHNOLOGIES USED

- OpenAI API (Chat Completion, Text-to-Speech, etc.)
- LangChain (langchain-community, langchain-experimental)
- Chromadb (Text Splitting, Vectorstores, Embeddings, etc.)
- Streamlit (for server-streaming)
- Fine-Tuning, Evaluation, Moderation, Hallucinations Prevention

MODERATION

- Evaluate Inputs, Moderation to create Responsible AI (no hate, self-harm, sexual content, and violence)
- Preventing Prompt Injection - users manipulating the AI system to bypasses intended instructions:
  - Using Delimiters and Clear Instructions in System
  - Using an Additional Prompt
- Few-shot Learning ⇒ The LLM model learns desired behavior by example

FEATURES

AI-Driven Multilingual Support

Speech Recognition

Dynamic Information Retrieval

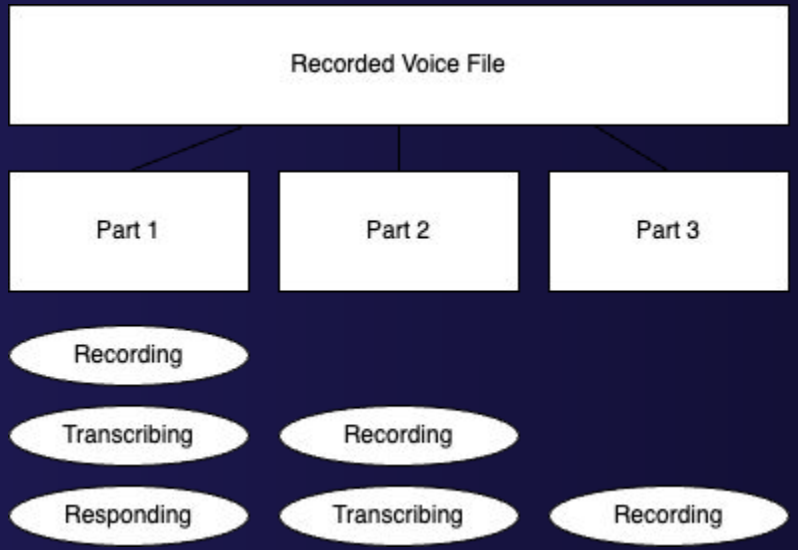
Universal Accessibility

Email Generation

Presented By-
Chenxin Cao

REAL-TIME SPEECH - RECOGNITION

Queue Streaming - Improve the responding speed



Presented By-
Chenxin Cao

SPEECH - RECOGNITION

OpenAI Whisper

- **Introduction to Whisper**: A speech recognition model developed by OpenAI.
- **Accuracy Across Languages:** Exceptional performance in diverse languages, making it ideal for multilingual support.
- **Robust in Noisy Environments:** Maintains accuracy even in challenging audio conditions.
- **Adaptability:** Easily integrates with existing systems, enhancing our chatbot responsiveness and efficiency.

Presented By-
Chenxin Cao

SPEECH - OUTPUT

OpenAI & Google TTS Integration

Natural and Expressive Speech: Both technologies provide high-quality, natural-sounding voice responses, enhancing user interaction.
Wide Language Support: Expansive language and dialect coverage ensures inclusivity for our diverse user base.

Presented By-
Chenxin Cao

FINE TUNING

- Fine-tuning is the process of customizing pre-trained AI models to adapt to specific tasks or domains.

- In the context of our school chatbot, fine-tuning allows us to tailor the AI model's responses to better suit the unique needs and requirements of our school community.

Presented By-
Prachi Sethi

FINE TUNING

Process Overview

- Data Collection
- Preprocessing
- Fine-Tuning (Training the model)
- Evaluation
- Iteration

Presented By-
Prachi Sethi

FINE TUNING

Benefits and Impacts

- Personalization
- Accuracy
- User Satisfaction
- Efficiency
- Continuous Improvement

Presented By-
Prachi Sethi

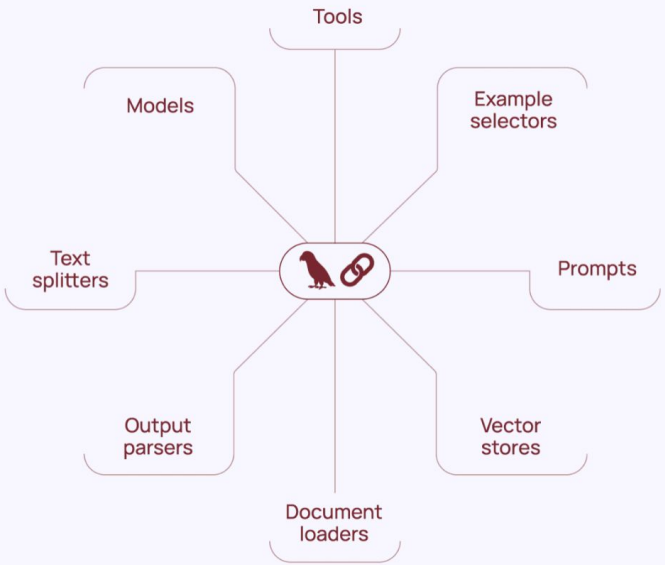
LANGCHAIN AND DOCUMENT LOADING

What is Langchain?

A complete set of interoperable and interchangeable building blocks

Leverage our comprehensive library of components that together make up sophisticated, end-to-end applications. Want to change your model? Future-proof your application by making vendor optionality part of your LLM infrastructure design.

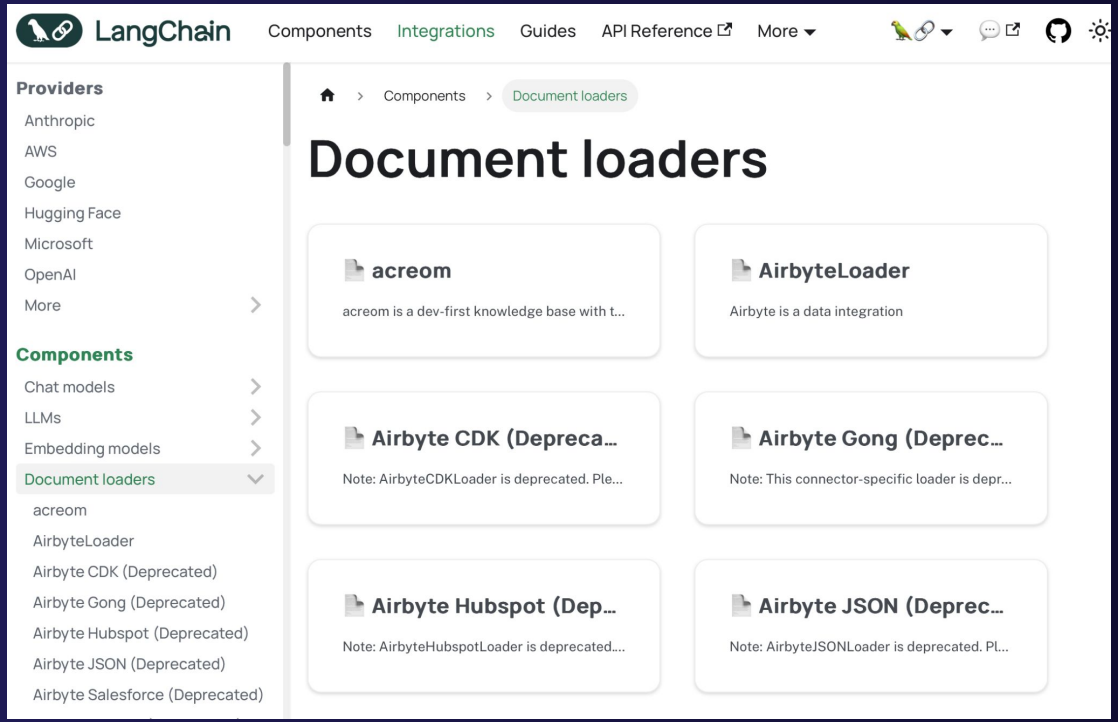
Go to Docs ↗



Presented By:
Chenxin Cao

LANGCHAIN AND DOCUMENT LOADING

Langchain offers a significant amount of document loaders

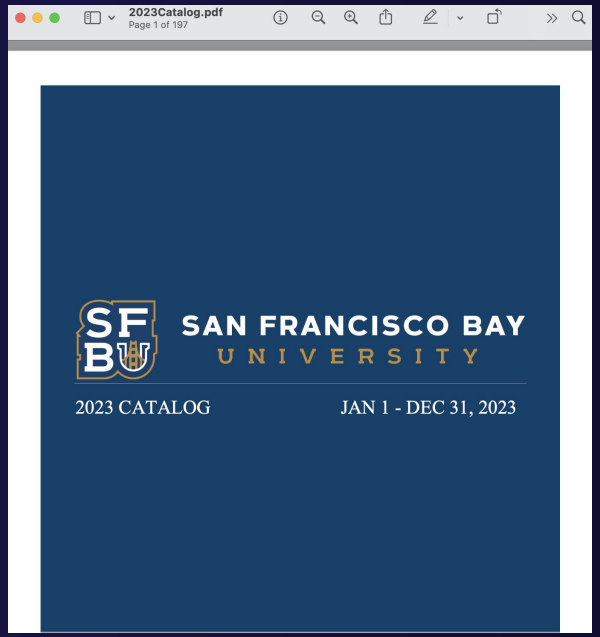


Presented By-
Chenxin Cao

LANGCHAIN AND DOCUMENT LOADING

PyPDF

Portable Document Format (PDF), standardized as ISO 32000, is a file format developed by Adobe in 1992 to present documents, including text formatting and images, in a manner independent of application software, hardware, and operating systems.



Our Primary Document: 2023 Catalog.pdf

Presented By-
Chenxin Cao

RAG

- **Definition**: Retrieval-Augmented Generation (RAG) is a method that combines a language model with a retriever that searches external knowledge sources to provide additional context for generating more accurate and relevant responses.

- **How It Works**:
  A retriever fetches relevant documents or data from external sources based on a query or prompt.
  The language model uses this retrieved information to generate a response.

- **Purpose**: Enhance the quality, relevance, and factual accuracy of responses by leveraging up-to-date information from external sources.

Presented By-
Prachi Sethi

RAG

- Improved Accuracy: Incorporates up-to-date information, leading to more accurate and fact-based responses.
- Contextual Responses: Uses additional context to provide more nuanced and context-aware answers.
- Reduced Hallucination: By grounding the language model's output in retrieved data, the risk of generating false or misleading information is minimized.
- Customizability: Can be fine-tuned to specific domains or use cases by focusing on relevant sources.

Presented By-
Prachi Sethi

SAFETY AND PRIVACY

- Prompt:
  - Injections
  - Proactive Injection Detection
- Prompt & Response:
  - Hallucination
  - Personally Identifiable Information (PII)
  - Toxicity
  - Sentiment analysis (Prevention from Negativity, Hate, Assault Speech)
  - Regexes (Regex pattern matching for sensitive information)

FUTURE SCOPE

- Enhanced Natural Language Understanding (NLU)
- Multimodal Support (voice input/output, image recognition, and video responses)
- Personalization and Context Awareness
- Integration with Existing Systems
- Feedback Mechanisms
- Language and Cultural Adaptability
- Integration with Social Media and Messaging Platforms

CONCLUSION & CHALLENGES

CHALLENGE 1: Which Model?

CHALLENGE 2: API Upgrades

Presented By-
Chenxin Cao

CONCLUSION

Why choose us?

Presented By-
Chenxin Cao

DEMO

Q & A

THANK YOU