SARCATHON

Pan-IIT Competition

Project: Smart FAQ Module for SARAS AI Institute

Pradipta Sundar Sahoo
Indian Institute of Technology (IIT) Roorkee
pradipta_ss@ch.iitr.ac.in

Agenda

- About the project
- Features
- Proposed Solution
- Results
- Access links
- Future Advancements
- Setup Instructions
- Usage
- File Structure

Tip: Use links to go to a different page inside your presentation.

How: Highlight text, click on the link symbol on the toolbar, and select the page in your presentation you want to connect.

About

- Goal: Develop a **Smart FAQ Module** for the **SARAS AI Institute** website that improves user experience by intelligently returning relevant FAQ entries based on user queries.
- Key Objectives:
  - Provide accurate and fast answers.
  - Ensure easy integration into the existing website.
  - Use open-source technologies for a cost-effective, scalable solution.

BACK TO AGENDA PAGE

Key **Features**

Relevance

Matches user queries to the most relevant FAQs using NLP techniques.

Latency

Optimized for quick response times even with large datasets.

User-Friendly

Clean, intuitive interface for easy interaction.

Scalability

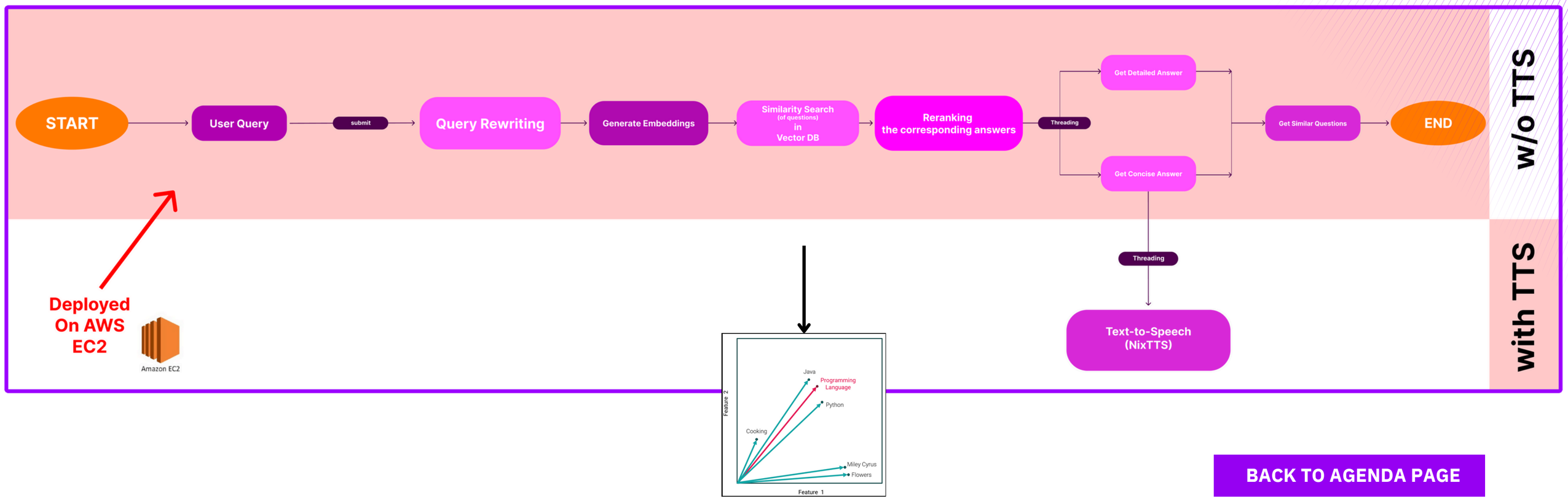
Handles large volumes of FAQs and multiple user queries.

Customizable

Easily adaptable to new FAQs and evolving user needs.

BACK TO AGENDA PAGE

Prosposed **Solution**



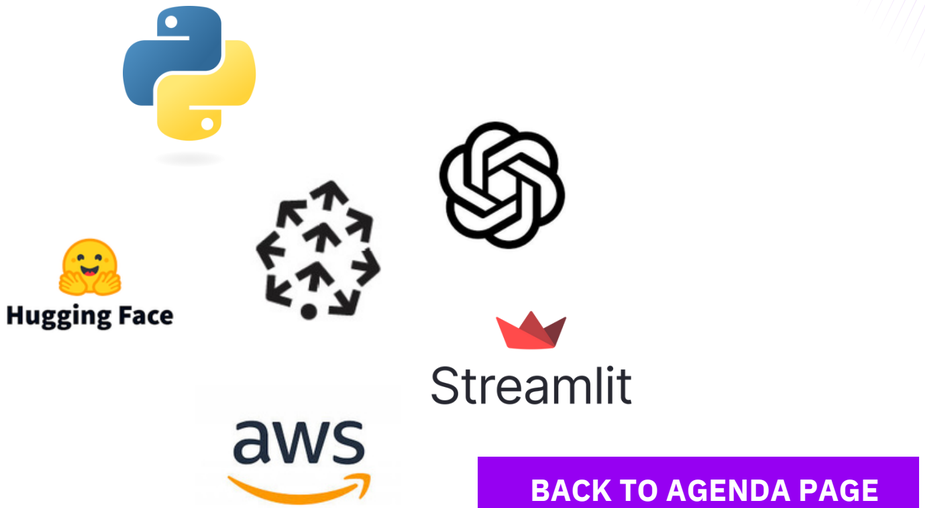
BACK TO AGENDA PAGE

Techniques **used**

- **Natural Language Understanding**: Uses OpenAI's language models to understand and interpret user queries.
- **Similarity Search:** Employs FAISS for fast and accurate similarity matching with the FAQ database.
- **Query Rewriting**: Enhances the clarity and relevance of user queries for better search results.
- **Concise & Detailed Responses**: Provides both quick and in-depth answers based on user preference.
- **Text-to-Speech (TTS) Option**: Converts the concise response to audio using NixTTS for added accessibility (optional)
- **User-Friendly Interface**: A clean, intuitive interface for easy interaction.
- **ADDED FEW MORE Questions: For robustness.**

Technology **used**

- **OpenAI**: For natural language processing and embeddings.
- **HuggingFace**: Have also implemented using hugging face.
- **Pinecone**: used it as vector DB.
- **FAISS**: For efficient similarity search on the FAQ database.
- **Streamlit**: For building a responsive web interface.
- **NixTTS**: For converting text responses to audio.
- **pyaudio & wave**: For audio playback in the TTS version.
- **Python**: Primary programming language for backend logic.
- **AWS**: Deployed it using AWS EC2.



BACK TO AGENDA PAGE

Why this **Architecture?**

Current model ⟶ **RAG** ✓

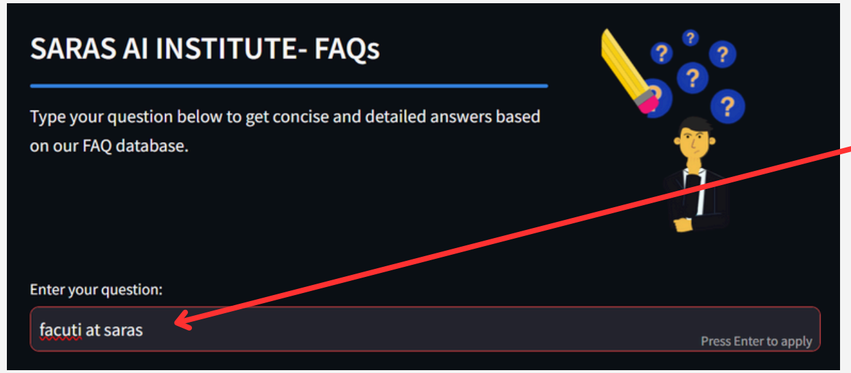
- **Real-time updates(dynamic RAG)**
- **Flexible query handling**
- **Economic**

vs

Fine-Tuning ✗

Results

Query



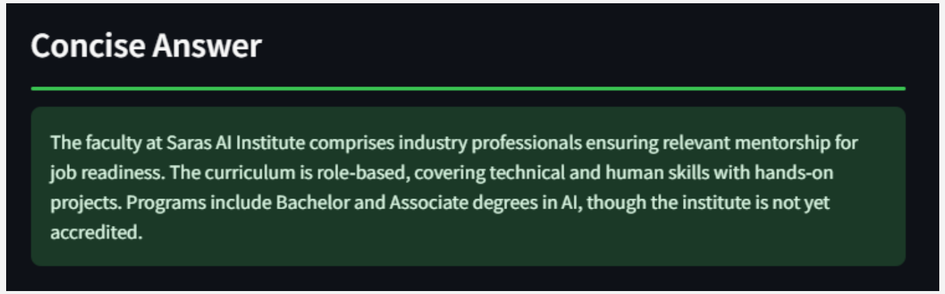
knowingly did a typo,

handled by query rewriting.

Rerank based on relevance score

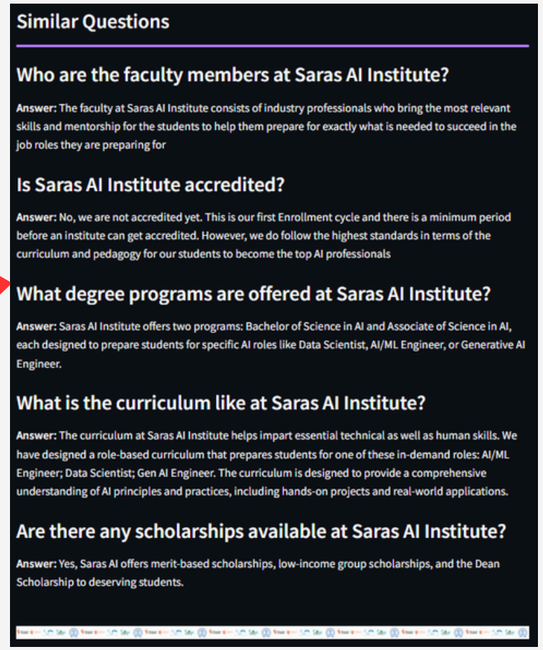
Output :

Concise Answer

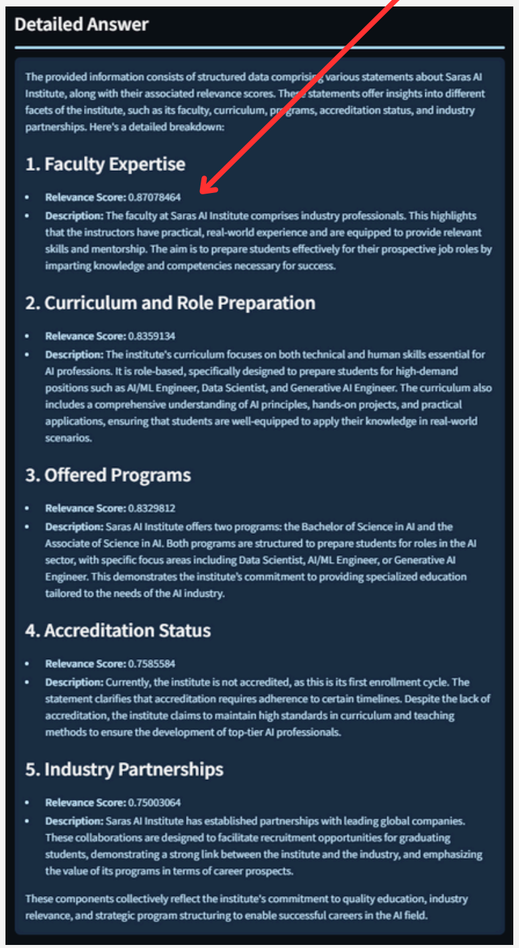


Similar Questions

Using vector search, found similar top-k questions



Detailed Answer



Access **Links**

Live Access Link:

SMART FAQ System- SARAS AI

Demo Video Link:

Demo Video Link

Github Link:

Github Repo Link

BACK TO AGENDA PAGE

Evaluation? RAGAS

Metrics	Scores
Faithfulness	0.98
Answer Relevancy	0.95
Context Precision	0.92
Context Recall	0.95
Context Entity Recall	0.95
Answer Similarity	0.96
Answer Correctness	0.94
Harmfulness	0.00

Future **Advancements**

Multimodal input and output:
have already been implemented.(**in repo**). But do we actually need it?

Multilingual:
knows how to implement. Could be useful !!!.

Faster Inference :
could use small llms.

API ⟶ Serverless

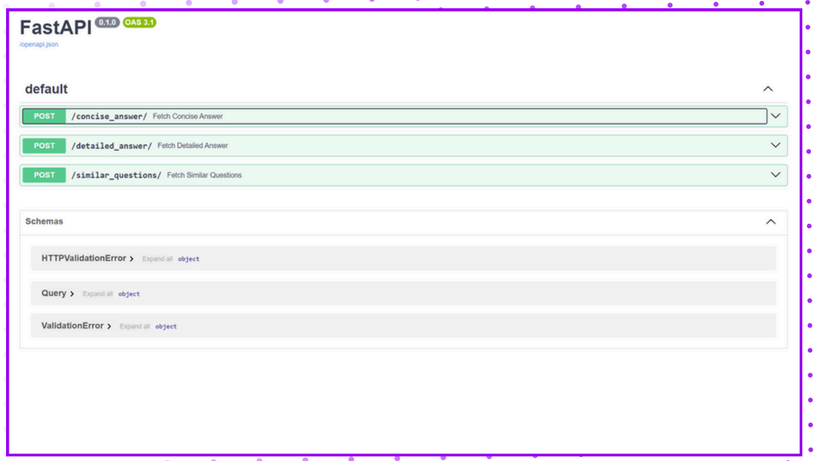
Loading Model ⟶ EC2

Scalability:
have already implemented a **FastAPI** code. It could be used for production. (**in repo**).

Relevance:
Reranked all results according to relevance. May a more robust way be done.

More robust:
I have already some extra info about SARAS AI. But more info can be fed.

Cost-effective:
Instead of EC2, can be deployed in serverless instances.
More intermediate stuffs can be removed to lower the tokens count.



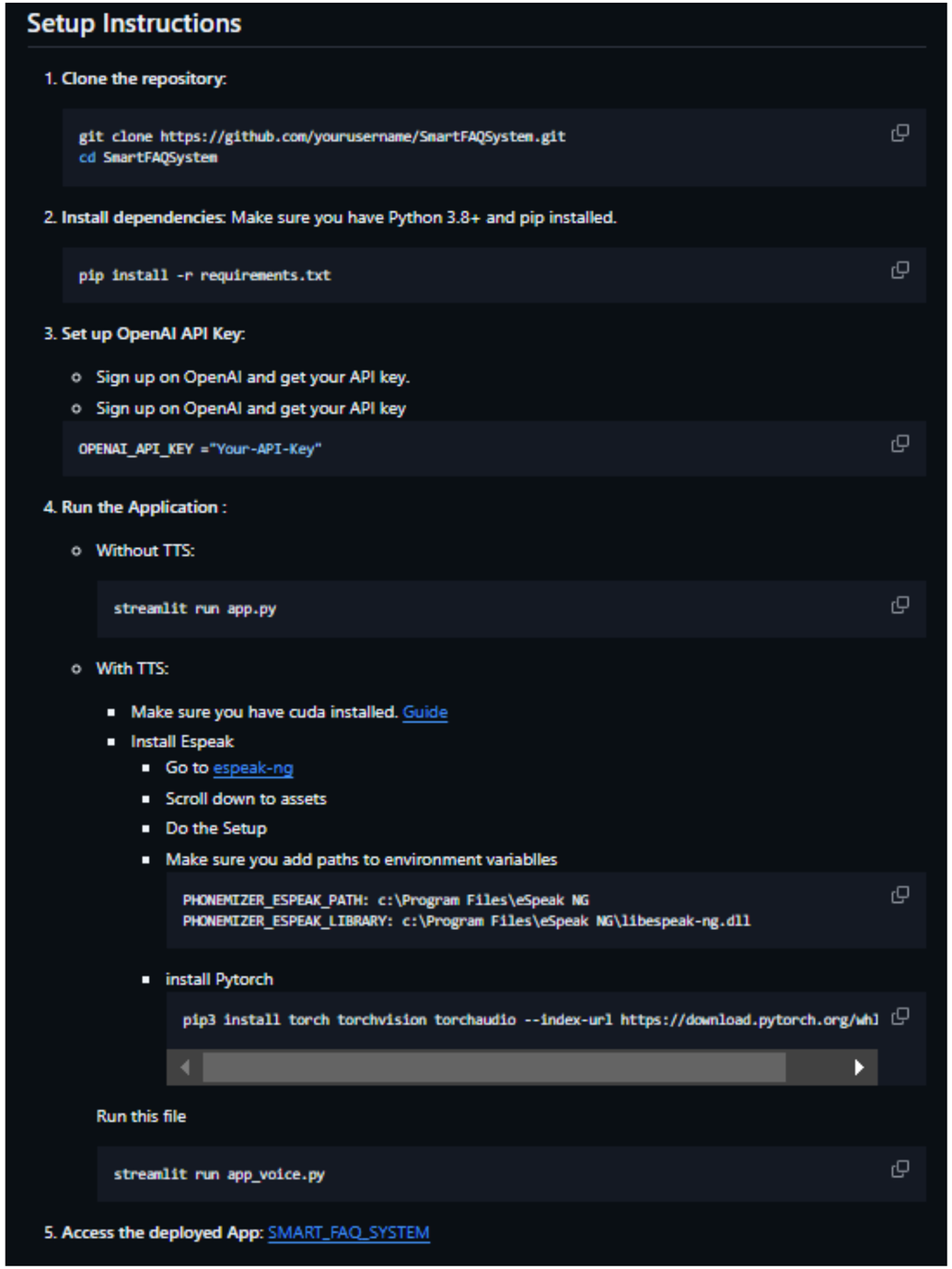
WHAT ELSE I HAVE TRIED?

- Tried with pinecone (fetching time was more)
- Tried Agentic Frameworks (Takes a lot of tokens and latency is high, but no such improvements)
- Gave multi-modal input i.e speech (found it useless for FAQ system)
- Can do multilingual.

Setup **Instructions**

Github Repo Link

BACK TO AGENDA PAGE



Setup Instructions

1. Clone the repository:

```
git clone https://github.com/yourusername/SmartFAQSystem.git
cd SmartFAQSystem
```

2. Install dependencies: Make sure you have Python 3.8+ and pip installed.

```
pip install -r requirements.txt
```

3. Set up OpenAI API Key:

- Sign up on OpenAI and get your API key.
- Sign up on OpenAI and get your API key

```
OPENAI_API_KEY ="Your-API-Key"
```

4. Run the Application :

- Without TTS:

```
streamlit run app.py
```

- With TTS:
  - Make sure you have cuda installed. Guide
  - Install Espeak
    - Go to espeak-ng
    - Scroll down to assets
    - Do the Setup
    - Make sure you add paths to environment variablles

```
PHONEMIZER_ESPEAK_PATH: c:\Program Files\eSpeak NG
PHONEMIZER_ESPEAK_LIBRARY: c:\Program Files\eSpeak NG\libespeak-ng.dll
```

    - install Pytorch

```
pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl
```

Run this file

```
streamlit run app_voice.py
```

5. Access the deployed App: SMART_FAQ_SYSTEM

Usage

Github Repo Link

Usage

1. **Enter Your Question:** Open the app in your browser and type your question in the input box.
2. **Click "Submit":** After entering your question, click the "Submit" button.
3. **View the Results:**
   - The app will process your question and display a concise answer and a detailed answer based on the FAQ database.
   - You'll also see a list of similar questions related to your query.
4. **Listen to the Answer (TTS Version):**
   - In the version with Text-to-Speech (TTS), the concise answer will be played as audio, providing an accessible option for auditory learners.
   - The audio playback is processed in chunks for efficiency and clarity.

File **Structure**

Github Repo Link

File Structure

```
SARCATHON/
│
├── app_voice.py            # Main app file with TTS feature
├── app.py          # Main app file without TTS
├── faqs.json                   # JSON file containing FAQ data
├── faq_index.faiss             # FAISS index file for similarity search
├── questions.npy               # Numpy file containing question embeddings
├── answers.npy                 # Numpy file containing answer embeddings
├── requirements.txt            # List of dependencies
├── .env                        # Environment variables file for API keys
└── README.md                   # Project documentation

└─fastapi_code.py     # Code for deploying , So that can be used for future advancement. [FAST API]
└─index_faq.py        # Code for creating index
└─ignore rest
```

BACK TO AGENDA PAGE

BACK TO AGENDA PAGE

Thank **You**

About **Me**

- **Avid Learner with strong problem solving skills.**
- **Keep tracks of recent GenAI Hacks**
- **Worked with multiple startups (healthtech,legaltech,fintech) build their MVP.**
- **Collaborated with cross-domain teams, showcasing good leadership and team management skills.**
- **Currently, focussing on AI Agents Architecture.**

Pradipta Sundar Sahoo
UG (III Year I Semester)
B.Tech. (Chemical Engineering)
Contact No: 7327873628
Email: pradipta_ss@ch.iitr.ac.in
Registration No: 22118054/2025

Indian Institute of Technology Roorkee

Education

Year	Degree/Examination	Institution/Board	CGPA/ Percentage
2024	B.Tech. 2nd Year	Indian Institute of Technology, Roorkee	8.271
2022	Intermediate (Class XII)	Shiv Jyoti Senior Secondary School , Kota	94.00 %
2020	Matriculate (Class X)	Vivekananda Shiksha Kendra , Bhubaneshwar	93.20 %

Experience

Co-founder | Lawcrats — August 2024 - October 2024
- Co-founded a legal tech AI startup, led a 12-member team, and secured support from Google, Microsoft, AWS, MongoDB; approved for NVIDIA Inception Program and Wadhwani Foundation cohort'24.
- Engaged with lawyers and advocates through meetings and discussions to align technology solutions with legal industry needs. I had to learn and implement a new technology.

Internships

AI Engineer intern | Anguliyam — August 2024 - October 2024
- I have developed end-to-end AI healthcare RAG applications for structured outputs, further incorporating data extraction, AI app development, and model deployment.
- Created conversation voice bot (s2s)using NixTTS and local llms.Reducing traditional latency from 32 s to 2s.

AI Engineer Intern | Tecosys (Acquired by Medynex) — June 2024 - August 2024
- Engineered and optimized AI RAG systems and tools, integrating Gen AI with Python to deliver scalable, high-performance applications, reducing costs and enhancing efficiency.
- Demonstrated creativity, problem-solving, and teamwork, consistently delivering innovative solutions under pressure with precision and attention to detail.

Projects

Praml An Automl | Indian Institute of Technology Roorkee — April 2024 - May 2024
- Created Praml, a Python Automated Machine Learning Library, to streamline building ML models by automating data preprocessing, model selection, hyperparameter tuning, and evaluation.
- Integrated diverse models for classification and regression with custom-made Tree-Parzen Estimator and hyperopt optimization.

Extreme low light denoiser | Indian Institute of Technology Roorkee — May 2024 - June 2024
- Developed an extreme low-light denoiser using multiple architectures including FFDNet, SwinIR, ResNet and UNet.
- ResNet emerged as the best-performing architecture with a PSNR of 25 after 40 epochs, significantly improving image quality in low-light conditions.

Monte Carlo Simulation Using Python | Indian Institute of Technology Roorkee — October 2023 - November 2023
- Implemented a Monte Carlo simulation in Python utilizing the Sharpe ratio to analyze and optimize investment strategies.
- Generated graphs to visualize simulated portfolio returns and risk-adjusted performance, providing insights into optimal asset allocation and risk management strategies.

Object detection using Opencv and Yolov8 | Indian Institute of Technology Roorkee — February 2024
- Developed an object detection system to identify and classify objects in images and video streams.
- Utilized OpenCV for image processing and integrated YOLOv8 for real-time detection.
- Enhanced accuracy and efficiency in surveillance, autonomous vehicles, and robotics applications.

Stock Sentiment Analysis Using Machine Learning | Indian Institute of Technology Roorkee — May 2024 - June 2024
- Developed a system to predict stock price movements by analyzing sentiment from financial news headlines using machine learning.
- Scraped financial news data using Beautiful Soup and visualized trading signals on stock price charts to enhance decision-making.

Skills

Computer languages: Python, C++, Js
Software Packages: PyTorch, TensorFlow, Linux, Git ,CI/CD, Docker, Kubernetes, PostgresSQL, MongoDB, OpenAI, Ollama, Langchain, Tableau, Airflow, REST APIs, GraphQL, Flask, FastAPI, CI/CD, AWS, Azure

Positions of Responsibility & Extra Curriculars

Delegate | FICCI India AI — October 2024
- Invited as a delegate to the FICCI India AI Conclave'24.
- Engaged in discussions with co-founders and heads from Google, AWS, and Panasonic on AI development initiatives.