VISMAY Jain

Contact Information:

📞 6351464527 | 📧 Vismay Jain | 📍 Surat, Gujarat, 395009

CAREER SUMMARY

AI and ML Engineer with a strong background in Data Science, specializing in the development and deployment of advanced machine learning models and AI-driven solutions. Experienced in collaborating with cross-functional teams to enhance operational efficiency and drive innovation. Proven track record of utilizing state-of-the-art technologies such as Groq API with LLaMA 370B, LangChain, and AWS to deliver robust and scalable solutions. Adept at voice assistant development, natural language processing, and implementing data synchronization systems. Seeking a position to contribute to cutting-edge projects and help technology groups achieve superior operational insight and performance.

WORK EXPERIENCE

Logictrix Infotech

AI/ML Intern **May 2023 - July 2023**

- Collaborated with Data Engineers and Web Development teams to support effective data analysis.
- Collected and resolved data discrepancies, ensuring data integrity for analysis.
- Identified and fine-tuned state-of-the-art (SOTA) pre-trained models for text processing.
- Conducted performance analysis and hyperparameter tuning to select the best model.
- Fine-tuned the ALBERT model on a custom email dataset for email classification.
- Worked with the product team to automate email classification using ALBERT and deployed the solution on AWS.

Commercient

AI/ML Intern **Dec 2023 - June 2024**

- Developed data synchronization DLL for syncing data between ODBC databases, QuickBooks, and Salesforce.
- Utilized state-of-the-art large language models (LLMs) for summarizing helpdesk tickets with over 200k input tokens using LangChain and multiple GPUs for inference. The LLM used was Mixtral 8x7B, which completed summarization in 3 days with 10% GPU usage each (4 GPUs).
- Upserted summarized data into Pinecone vector database using LangChain and developed a RAG (Retrieval-Augmented Generation) chat assistant to handle user requests.
- Employed OpenAI embeddings and LLMs, specifically LLaMA 70B via Groq API, to enhance response quality and speed.
- Optimized the company's existing RAG solution, reducing response time from 20 seconds to 5 seconds.
- Collected data from YouTube and Gmail, and performed OCR for additional data sources.

Projects

- **Voice Assistant (NLP):**
  - Developed a voice-controlled AI assistant using Python, integrating multiple APIs and libraries for diverse functionalities.
  - Implemented wake word detection using Porcupine, enabling activation with the phrase "snowman."
  - Utilized Groq API with LLaMA 370B for natural language processing to generate interactive and context-aware responses.
  - Integrated various functionalities, including:
  - Opening applications and websites
  - Playing Spotify songs via Spotify API
  - Generating images from descriptions using a stable diffusion pipeline
  - Sending WhatsApp messages through voice commands
  - Ensured accessibility by enabling voice responses with the pyttsx3 library, making the assistant both voice-activated and voice-responsive.
  - Employed LangChain tools and agents to handle diverse user requests via voice input, enhancing versatility.

- Used PlayHT API for realistic female voice output, providing a natural voice assistant experience.
- Leveraged Groq API for the LLM agent, integrating multiple data sources to assist users with urgent help, debate practice, and daily problem-solving.

- Mail Classification System Using BERT:

- Developed a Flask-based web application for email management and classification.
- Implemented MySQL database integration to store email metadata and classification results.
- Utilized the Transformers library to fine-tune a pre-trained Albert model for email classification.
- Set up a periodic email fetching mechanism using IMAP to process incoming emails.
- Created a responsive web interface for users to view, classify, and search emails.
- Implemented a feature to reply to emails, automatically classifying and storing the reply.

-AI Posture Correction Assistant (APCA):

- AI Posture Correction Assistant (APCA) is a real-time posture analysis and correction system.
- Utilizes the MediaPipe Pose model to detect and analyze the user's body posture through a webcam feed.
- Provides feedback and recommendations for improving posture based on real-time measurements of shoulder width, arm length, and leg length.
- Incorporates a voice-activated chatbot named Suzuki to interact with users and offer guidance.
- Supports voice commands for initiating posture correction and changing the chatbot's voice.
- Enhances user experience with speech-to-text and text-to-speech capabilities.
- Allows users to receive the current time and polite responses for expressions of gratitude.
- Empowers users to initiate posture correction sessions for self-improvement.

EDUCATION

Uka Tarsadia University
Bachelor of Technology in Artificial Intelligence and Data Science **June 2024**

- CGPA 8.77/10.0

TECHNICAL SKILLS

Skills	Category
Python	Programming Language
TensorFlow	Machine Learning Framework
PyTorch	Machine Learning Framework
Deep Learning	Artificial Intelligence
C#	Programming Language
SQL	Database Management
RAG	Machine Learning Technique
Vector Database	Text Vector Database Management
LLM	Artificial Intelligence
Generative AI	Artificial Intelligence

AWS SageMaker	Cloud Service
AWS Lambda	Cloud Service
AWS BedRock	Cloud Service
Data Science	Field of Study
Computer Vision	Artificial Intelligence

Certificates

- **Amazon Machine Learning Specialty Certified**
  by AWS Credential ID (0156d1f6a8d544829a242df102fc3bd)
- **Amazon Web Services Cloud Practitioner**
  by AWS
- **Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning.**
  by Coursera
- **Data Analysis with Python**
  by Coursera
- **IBM Data Analysis**
  by IBM
- **Tweet Emotion Recognition with TensorFlow.**
  by Coursera
- **Introduction to Generative AI**
  by Google CLoud