VISMAY Jain

Contact Information:

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

CAREER SUMMARY

AI/ML Engineer with extensive experience in designing, developing, and deploying machine learning models and AI-driven solutions. Proficient in state-of-the-art technologies such as **Generative AI**, **Large Language Models (LLMs)**, and **LangChain**. Expertise in building scalable AI solutions, including chatbots, data synchronization systems, and natural language processing (NLP) tools. Adept at integrating AI systems into production environments, optimizing performance, and leveraging **multi-GPU systems** for high-efficiency inferencing. Seeking to contribute to innovative projects in a challenging environment.

WORK EXPERIENCE

Commercient
AI/ML Engineer
June 2024 – Present (Full-Time)

- **Developed a sales onboarding chatbot using LangGraph and Python, integrating it with C# APIs for seamless deployment.**
- **Spearheaded the use of Generative AI for creating advanced NLP models, leveraging LangChain and LLMs for contextual data handling.**
- **Managed multi-server, multi-GPU inferencing with LLaMA 3.1 using vLLM, achieving optimal load balancing for large-scale AI deployments.**
- **Enhanced chatbot capabilities on platforms like Zoom and Slack, providing robust integration and real-time responses.**

AI/ML Intern
Dec 2023 – June 2024

- **Designed and implemented a data synchronization DLL for seamless interaction between ODBC databases, QuickBooks, and Salesforce.**
- **Processed over 200k input tokens using LangChain and Mixtral 8x7B LLMs across multiple GPUs, summarizing helpdesk tickets with a 10% GPU utilization rate.**
- **Engineered a RAG (Retrieval-Augmented Generation) system leveraging Pinecone vector databases and OpenAI embeddings, reducing response times from 20 seconds to 5 seconds.**
- **Integrated external data sources, including OCR, YouTube, and Gmail, for comprehensive data aggregation.**

Logictrix Infotech
AI/ML Intern
May 2023 – July 2023

- **Collaborated with data engineering and web teams to streamline data pipelines and ensure high-quality data analysis.**
- **Fine-tuned ALBERT on custom datasets for email classification and deployed the solution on AWS for automated workflows.**
- **Conducted performance analysis and hyperparameter tuning to achieve optimal model accuracy and efficiency.**

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

Bachelor of Technology in Artificial Intelligence and Data Science
Uka Tarsadia University
June 2024 | CGPA: 8.77/10

TECHNICAL SKILLS

Programming: Python, C#, SQL,NextJS
AI Frameworks: TensorFlow, PyTorch
Cloud Services: AWS SageMaker, AWS Lambda, AWS BedRock
AI Techniques: Generative AI, LLMs, RAG, NLP
Databases: Pinecone Vector Database
Tools: LangChain, LangGraph, vLLM, OCR

Certificates

- **Amazon Machine Learning Specialty Certified (AWS)**
- **AWS Cloud Practitioner (AWS)**
- **Introduction to TensorFlow (Coursera)**
- **Introduction to Generative AI (Google Cloud)**
- **Tweet Emotion Recognition with TensorFlow (Coursera)**