Harsh Vardhan

Generative AI Engineer

harsh.vardhan7695@gmail.com 9583282224 Bangalore Linkedin Medium

AI Enginner with experience in building real-time AI solutions, intelligent autmation pipeine and various end-to-end GenAI solutions. Also with have experienced in NLP, Python, SQL and various machine learning and Deep learning techniques, with a proven track record of deploying models in production environments, and expertise in data-driven decision-making.

PROFESSIONAL EXPERIENCE

Generative AI Engineer, *Genpact* 07/2024 – present

- **Real-Time AI Agent for Customer Retention (In Progress):** Engineering a low-latency, real-time AI agent system to handle customer complaints autonomously using speech-to-text (STT), multi-agent orchestration (LangGraph), LLM reasoning, and text-to-speech (TTS).
- **Audio2Action :** Designed and deployed a robust AI system to convert enterprise audio content (scrum calls, webinars, knowledge sessions) into structured summaries, insights, and task assignments. Leveraged **multi-agent orchestration frameworks (CrewAI, Autogen,Langgraph)**, and Azure Services like Azure model from Azure AI foundry(Gpt 4o), Azure Real time diarizer for Transcription and diarization, and Azure Speech synthesis for generating final audio digest.Achieved high accuracy with **WER as low as 6.2%** and **MOS up to 4.7/5**
- **Retrieval-Augmented Generation (RAG) Pipelines – Multimodal & Comparative Analysis:** Designed and implemented multimodal RAG pipelines leveraging LlamaParse, ColBERT-Qwen2, and ColBERT-Pali and many other LLM model for diverse document types (Excel, PDF, PPT, DocX). Created Q&A chatbots specialized on domain-specific use cases such as ProcessMap and Engineering Diagram data interpretation.
- **Order Management System –** Developed end-to-end pipeline with finalized high-accuracy (>92% accuracy) for the Order Management system, enhancing automated order processing.

Senior Software Engineer, *First American India* 05/2022 – present

- **FASTSearch**
  Technology & Packages Used – Python, Langchain, Claude, Pinecone,AWS Bedrock
  - Designed and implemented a data pre-processing pipeline to convert,extract and standardize data from diverse document formats.
  - Developed a vector indexing and similarity search framework using PineconeDB to facilitate the identification of documents closely matching the query vectors.
  - Created a QnA RAG application using Langchain framework to help agent with their query in processing the files.
- **Home Warranty Sentiment Analysis**
  Technology & Packages Used – Python, HugginFace BERT, Power BI.
  - Implemented a system to find out the actual sentiment of the user reviews provided on Home Warranty Website.
  - A sentiment Analysis model was built using BERT to identify the sentiment of every sentence from the review paragraph.
  - The Paragraph was split into sentences using sentence tokenizer and then sentiment for every review was provided.
  - Topic modelling was also done with sentiment for every review to find out if a customer is talking about which aspects of the service like Quality, Design, etc.
  - PowerBI Dashboard was built to showcase various metrics and KPI's to the Stake Holder like number of positive and negative sentiments, Top keywords for Positive Sentiments, Top Keywords for Negative Sentiments.
  - Word cloud was also shown in dashboard to provide top keywords about every topic extracted.
- **AUTO ADO**
  Technology & Packages Used – Transformer, BERT, HuggingFace
  - Developed a solution which Summarizing meeting outcomes.
  - Implemented sentiment Analaysis of all the mail which were involved during a particular conversation.
  - Design readily available course of action required with reduced time and less human intervention.
- **Title Insurance Price Prediction and Recommendation**
  Technology & Packages Used – Python, Random Forest(Scikit learn), Pandas, EDA, AWS(S3,ECR,Sagemaker,CloudWatch, MLFlow)
  - Created an End-to-End system which predicts the title insurance price of the real estate properties.
  - Predicted the title insurance price of property on the basis of the input given by the user with 87% accuracy.
  - Recommended 5 more properties using Content Based filtering.

- Implemented Analytics module which does Exploratory Data Analysis.
- Model deployment and monitoring solution was done using AWS.

System Engineer, *Infosys* 2018 – 2022

- **Mobile Ads Click Through Rate Prediction**.
  Technology & Packages Used – Xgboost Model, Power BI,Pandas
  - This system helped the client to bring more customers to a specific website through the marketing campaigns.
  - The system is able to increase the number of customers for an agricultural tourism website.
  - In this, Feature Engineering was done on the given data to remove the anomalies and filter the data in the proper format to apply for specific ML Algorithms.
  - The Prediction was done on a specific model to find the probability that the user will visit the site on clicking the Ad.
  - Here a LightGBM model and Xgboost Model were used for training on a given dataset.
  - Customized reports and metrics dashboard were built using Power BI to provide useful insights to the client about the website on daily basis.
- **Human Resources HR Analytics on Employee Attrition.**
  Technology & Packages Used – XGBClassifier, SMOTE
  - A system was developed which helped to save the money and time of the client for re-recruiting other employees.
  - The system was based on a dataset that consists of various characteristics of employees and it was labelled whether the employees are still in the company, or they have gone to work somewhere else.
  - At first, the baseline model was trained and an F1-score of 50.5% and Recall = 48.9% was achieved.
  - Then after feature selection and feature engineering, F1-score = 49.6% and recall = 61.7% was achieved but this was low because the data was highly imbalanced.
  - After applying Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class, some improvement in both F1-score and recall was observed. i.e., F1-score = 55.7% & recall = 68.1%. So, with this accuracy 82.7% was achieved.

SKILLS

GenAI
LLM(OpenAI,Llama, Databricks Models etc), RAG,Multimodal(ColPali,ColQwen etc),AI Agents framework(CrewAI,Langgraph,Autogen)HuggingFace , Langchain

Data Science Toolkit
Pandas, Numpy, Matplotlib, Seaborn, Stats Model, Scikit learn, SciPy, NLTK, Spacy, OCR, OpenCV.

Deep Learning Frameworks
Keras,Tensorflow, Pytorch

Deep Learning Algorithms
ANN, RNN, LSTM, GRU, CNN , BERT

Machine Learning
Decision Trees, Random Forest, Gradient Boosting, Linear Regression, Logistic regression, K-Nearest Neighbors, Clustering

Database Management
SQL Server, MySQL, Postgres SQL, Aurora, Vector Database(Cassandra, PineCone)

EDUCATION

Bachelor of Technology in Computer Science, 2014 – 2018 | Berhampur(Odisha)
National Institute of Science and Technology

PERSONAL PROJECTS

Dietplanner present
An AI powered personal Nutriotionist. DietPlanner crafts personalized meal plans based on your health goals, medical conditions, and preferences using multiple AI Agents that does the heavy lifting for you, so that you can do heavy lifting in the workout.

Q&A Chatbot from PDF
Technology and Packages Used- OpenAI, Langchain, PineCone,OpenAI Embeddings, Streamlit
- Developed a Q&A Application in which user can upload multiple pdf documents upto size(200mb) and then can ask question based on the context.
- Designed and implemented a data pre-processing pipeline to convert, extract, and standardize data from diverse document formats (PDF, Excel/CSV, DOCX) using tools such as docx2pdf, PyPDF and pandas, facilitating seamless input to GPT models.
- Employed Pinecone vector DB to store and manage extracted data and document embeddings(using OpenAI Embeddings), enabling effective vector similarity searches and efficient data retrieval.