Rohith Perumandla

Chicago,IL | rohith.perumandla.12@gmail.com | +17739432531 | www.linkedin.com/in/perumandla-rohith/ | github.com/Rohith1-p

SUMMARY

I'm an AI Engineer with 4+ years of hands-on experience building scalable machine learning and NLP systems powered by large language models. I've designed and implemented end-to-end conversational agents using techniques NLP techniques. I'm passionate about crafting robust dialogue systems that deliver natural, engaging interactions. I've successfully deployed these solutions on cloud platforms—ensuring they're reliable, scalable, and easy to integrate.

EDUCATION

DePaul University, Chicago, IL

Masters in Data Science • *June 2025*

Concentration: Computational methods - GPA: 4.0/4.0

- Coursework: Python Programming, Data Analysis Statistics and Regression, Fundamentals of Data Science, Digital Health, Advance and Programming in Machine Learning, Big Data Mining, Health Data Science, Advance Machine Learning, Deep Learning and Neural Networks
- Graduate Teaching Assistant for Advance Machine Learning, Health in Data Science and Topics in Digital Health

Rajiv Gandhi University of Knowledge Technologies, IIIT Basar

Bachelors in Electronics and Communication Engineering • *May 2022*

Linear Algebra, Calculus, Probability and Stochastic Process, Pattern Recognition, Deep Learning.

EXPERIENCE

DePaul University, ***Graduate Research Assistant – AI Engineer***

Sep 2023 - Present

- Designed and fine-tuned Large Languge Models based conversational agents (Llama3, Phi-3, BERT, T5) using PyTorch to assess cognitive decline in dementia patients, contributing to healthcare advancements.
- Implemented machine learning algorithms for biomarker extraction from multimodal inputs (speech, text, audio) to enhance diagnostic accuracy.
- Developed and deployed a scalable system on Google Cloud Platform (GCP) using Nginx, Docker, and WebSockets, enabling real-world testing, serving crucial Medicine applications.
- Optimized system architecture, reducing response time by 75%, significantly enhancing real-time patient engagement and interaction.
- Reduced GPU computational costs by 60% through model optimization, efficient inference strategies, and resource allocation improvements.
- Engineered linguistic biomarker extraction modules using TensorFlow, PyTorch, Hugging Face Transformers, scikit-learn.
- Monitored performance metrics and compliance with regulatory standards in a production environment, ensuring workflow efficiency and documentation of results.

Setuserv, ***Data Scientist – Customer Management***

Jan 2022 - Jul 2023

- Built and deployed an end-to-end scalable data science and machine learning system for extracting product reviews, ratings, and product details from e-commerce websites.
- Designed and automated MLOps data pipelines, reducing data processing time by 50% and improving workflow efficiency. Optimized data storage in relational databases
- Performed advanced data preprocessing, including data cleaning, stopword removal, and symbol elimination. Conducted exploratory data analysis (EDA) and feature engineering, leading to 15% improved sentiment classification accuracy in text analytics and NLP models.
- Decreased all pilot projects delivery time by 70% by NLP, and DL models like BERT, Name Entity Recognition (NER), Zero shot classification, and Word2Vec ML models from Hugging Face, Spacy Utilized: Huggingface Transformers.
- Built and managed MLOps including the core ML server REST API, pivotal in the firm's production pipeline, integrating ML models for large-scale client solutions and handling over half a million data points per day enabling the company to handle enterprise clients. Used: Python and Django.
- Improved the firm's data pipeline and data quality checks efficiency by integrating Google Sheets with backend servers. Leveraged: Python, Google Appscript, Django, sklearn, and PyTorch.

LICENSES & CERTIFICATIONS

Machine Learning Specialization
Stanford University • Issued Nov 2022

Python 3 Programming Specialization
University of Michigan • Issued Oct 2022

Introduction to Statistics
Stanford University • Issued Oct 2022

Unsupervised Learning, Recommenders, Reinforcement Learning
Issued Nov 2022

SKILLS

Database & Programming: OOPs, Python, R, Java, Matlab, Data Structures, SQLite, MongoDB, HTML, CSS, Javascript, Appscript, Arduino.
Machine Learning Frameworks: Numpy, Pandas, PyTorch, Keras, TensorFlow, Matplotlib, Seaborn, Supervised learning, Unsupervised Learning, Deep Learning, Clustering, RAGs. Agentic AI, Conda
Cloud & Software Tools: Amazon web services (AWS), Google Cloud (GCP), Django, Git, Kafka, Docker Containerization, Databricks.
AI Tools: Prompt Engineering, LangChain framework
Analytics tools: Proficient in Excel, Tableau

PROJECTS

Fine-tuned LLMs for Conversational AI Agents with RAG:
Developed an advanced conversational system using open-source Llama3 and Langchain RAG focusing on personalized dialogue generation. Employed state-of-the-art LLM techniques to optimize model performance, reducing GPU cost and latency and fine-tuning time while improving the system's generalizability to diverse patient interactions.

Bank Loan Case Prediction:
Developed a predictive model using historical bank loan datasets to assess loan repayment likelihood for better Financial Services; leveraged machine learning algorithms including Logistic Regression, SVM, Random Forest, Decision Tree, and K-Nearest Neighbors. Optimized model performance and evaluation, achieving a 93% accuracy rate, significantly enhancing data-driven lending decisions.

Brain Tumor Classification and Segmentation using Deep Learning:
Implemented deep learning models utilizing ResNet50 and convolutional neural network (CNN) architectures to accurately classify and segment brain tumors from medical images, achieving a segmentation score of 0.87 and a classification accuracy of 93%. Processed over 1,000 MRI images with CNNs, enhancing model performance through advanced data augmentation and transfer learning techniques.

Book Recommendation System using PySpark & SQL :
Built a scalable recommendation system processing 5M+ ratings using ALS collaborative filtering in PySpark. Designed an AWS based batch pipeline with S3 bucket, Athena SQL queries for accessing data, and Step Functions for automated ingestion, transformation, and model retraining. Achieved RMSE of 0.95 in validation, optimizing book recommendations with future scope for sentiment analysis & real-time deployment.