Abhishek Ranjan Prusty

+91-7077354197 | aviranjan444@gmail.com | Leetcode | Github | LinkedIn | Kaggle

Summary

Dynamic and highly motivated ML Engineer with hands-on experience in developing and deploying production-grade LLM systems and GenAI applications. Proficient in designing scalable AI solutions using RAG frameworks, MLOps pipelines, and classical ML algorithms with fair expertise in NLP, data-driven problem-solving, and software engineering first principles.

Experience

ML Engineer June 2024 – Present
Sprouts.ai *Remote*

- Working on building a standalone multimodal outbound sales agent.
- **AI Signals**: Engineered scalable real-time **RAG enrichment pipeline** for **open-domain question answering** about companies by retrieving search results, LinkedIn + Apollo data, & scraped website data.
- Increased the API throughput to approx. **3000 signals/min** and reliability by integrating asynchronous querying to LLM APIs with fallbacks via **LangChain**, & load balancing across multiple deployments (e.g. Azure/Vertex AI), with **usage based async** routing strategy.
- **Hyper-Personalized Messaging (HPM)**: Built a system to empower SDRs & BDRs for generating AI-driven personalized LinkedIn messages & email using contact and company information from multiple sources to produce variations of emails with different ice breakers and value propositions tailored to the recipients profile.
- Implemented **Mixture of Agents, context compression & asynchronous batch processing** for message generation, reaching a throughput of nearly 1000 messages/min and improving email engagement by **35%**.
- Optimized response quality & consistency by implementing **guardrails** on LLM generated response; reduced the latency by integrating **semantic caching/context caching** with efficient prompt engineering techniques cutting operational costs by **36%**.

Data Science Intern December 2023 – May 2024
Sprouts.ai *Remote*

- Productionized **Sprouts Cataloger** (an internal tool to standardize & catalog buyer persona attributes, specifically Job Titles & Seniority / Designation) using **PySpark**, refactored code for 3x faster processing, and standardized buyer persona datasets with enhanced taxonomy for improved search and filtering.
- Enhanced job title attribute extraction with **NLP**, **RegEx**, improving standardized format success by **35%**, and boosting attributes' coverage by approx. **40%** via reverse engineering and optimized lookup tables.

ML Engineer Intern Sept 2023 – Jan 2024
Kusho.ai *Remote*

- Worked on building an **AI agent for Autonomous API testing** using OpenAPI spec, leveraging **prompt chaining**, **RAG workflows**, and fine-tuned LLMs **(Llama-2-7b, Mixtral-8x7B)** via **PEFT** and **QLoRA**, achieving a 40% improvement in performance.
- Developed an internal tool for API test case validation with **Few-Shot Learning** and **Chain of Thought**, increasing accuracy by 30%.

Projects

Agentic Code Review System: Developed an AI-powered GitHub Action that automatically reviews pull requests and labels issues using an LLM hosted on **Azure OpenAI & the RoBERTa** model

AI Powered Company Culture Analytics Tool: Developed a system that aims to provide insights & analysis reports about a company's culture by analyzing multiple textual data sources with **RAG** & NLP algorithms.

Technical Skills and Coursework

Languages: Python, C/C++, SQL
Frameworks: PyTorch, LlamaIndex, TensorFlow, LangChain, scikit-learn, PySpark, FastAPI, Pandas, NumPy, XGBoost, HuggingFace, NLTK, LangGraph
Tools: Vertex AI, Databricks, Redis, Elastic Search, Milvus, Docker, Kubernetes, Git, Linux

Education

National Institute of Technology, Rourkela Odisha, India
B.Tech in Computer Science and Engineering, CGPA:8.06 *Oct. 2020 – May 2024*
Coursework: Data Structures and Algorithms Analysis, Machine Learning, Operating Systems, NLP with Deep Learning, LangChain & Vector Databases in Production