VIDYASAGAR KARUTURI

6309807183 ◇ Andra Pradesh, India
karuturividyasagar@gmail.com ◇ LinkedIn ◇ Portfolio ◇ Medium ◇ GitHub

PROFESSIONAL SUMMARY

GenAI Engineer and Data Scientist with a proven track record in designing, developing, and deploying end-to-end AI/ML solutions at scale. Deep expertise in Large Language Models. Skilled in model training, fine-tuning, and evaluation, with proficiency in deep learning frameworks like TensorFlow and PyTorch, and strong command of MLOps principles. Demonstrated success in building scalable machine learning pipelines, deploying production-grade AI systems and leveraging cloud platform like AWS for real-time AI applications. Seeking a challenging and rewarding role at a leading technology company where I can contribute to innovative AI products and advancements.

EDUCATION

Bachelors of Technology in Computer Science and Engineering (Specialization: Data Analytics),
Vellore Institute of Technology University, Amaravati 2019 – 2023
CGPA: 8.05 / 10
Relevant Coursework: Linear algebra, Object Oriented Programming, Databases, Discrete Mathematics, statistics, Operating Systems, Computer Networks, Machine Learning, Data Mining, Cloud computing

Intermediate Education (Science Stream),
Narayana Junior College, Vijayawada 2017 – 2019
CGPA: 9.5 / 10

Secondary Education (10th Grade),
St. Claret EM High School 2017
CGPA: 9.7 / 10

CORE TECHNICAL SKILLS

Programming Languages & Frameworks: Python (NumPy, Pandas, Scikit-learn, Keras, matplotlib), SQL
Databases: MongoDB, MySQL
Machine Learning: Supervised Learning (Decision Trees, Random Forests), Unsupervised Learning, Data Modeling & Evaluation, Preprocessing & Postprocessing, Model Optimization & Performance Tuning.
Rest-API: FastApi , Flask
Deep Learning: Pytorch, Tensorflow
Natural Language Processing: Classification, Summarization, Text Generation, Name Entity Recognition, Transformers, Retrieval Augmented Generation, ReAct, Retrieval, AgenticRAG, Langchain, Llamaindex, FastMCP, Ollama, Groq, PydanticAI
LLM Experience: Meta LLaMA 2, Google Gemma, OpenAI GPT-4, Anthropic Claude, MCP, CrewAI, Google ADK, Google Vertexai
Cloud & MLOps: Amazon Web Services (EC2, S3, IAM, Lambda, RDS, Sagemaker),Azure, Terraform, Docker, Kubernetes, CI/CD via Jenkins, Git, Postman
Tools: Visual Studio code, Cursor, Jupyter Notebook, CVAT, PuTTY, Tableau, Langfuse, Mlflow, Kibana

PROFESSIONAL EXPERIENCE

Data Scientist Jun 2023 – Present
Intellect Design Arena Pvt. Ltd., Chennai

- Designed and scaled a complex GenAI platform for leading financial firms to process, extract, and classify millions of diverse financial documents, with real-time data streams and scalable knowledge bases, achieving latency under 200ms, 97.9% uptime, and seamless horizontal scaling.

- Developed and integrated an MVP on Reasoning and Action Agents, Chain-of-Thought (CoT) and Tree-of-Thought (ToT) using Langchainacquisition, feature engineering, model development, evaluation, and cloud deployment using PyTorch and TensorFlow.
- Built LLM-based AI Guardrails for Prompt Injection Detection, Personally Identifiable Information (PII) masking, and Toxic Content Filtering using Meta LLaMA 2 70B, GPT-4o, and Gemma 2B with SETFIT and AWS Comprehend which safeguards 80% of attempts
- Integrated Smart Retrieval-Augmented Generation (RAG) with short term and long term memory using Redis and Cosmos DB to improve contextual awareness and user experience in financial service dialogues, while also implementing Cache Augmented Generation (CAG) to enhance response speed and reduce latency by preloading knowledge into the model's context window
- Designed and implemented an automated evaluation module to assess LLM and LMM-generated responses across Correctness, Faithfulness, and Relevancy metrics, leveraging GPT-4 and GPT-4o-based evaluators (LLM-as-aJudge and ROUGE-L, fuzzy JSON, exact match scoring, and accuracy metrics) for comprehensive quality assessment in RAG workflows.

Cloud Intern Feb 2023 – Apr 2023
LTI Mindtree

- Automated Infrastructure as Code (IaC) solutions using Terraform for AWS resource provisioning (EC2, S3, IAM, Lambda, RDS).
- Implemented security compliance and cost optimization strategies for cloud-based deployments.

PROJECTS

AI Guardrails System
Implemented robust guardrails using SETFIT and LLMs to detect prompt injection, PII exposure, and harmful content. Applied synthetic data generation via Gemma 2B for anonymization.

Citation Attribution Engine
Engineered citation backfilling to validate and attribute references used in generative responses, increasing more transparency and granularity.

LLM Evaluation Framework
Developed metrics-based evaluation pipeline for assessing benchmarking agent outputs using Judge, ROUGE-L, fuzzy matching, and JSON schema validation.

Automated Document Intelligence
Created a document classification and extraction model based on LayoutLM for automating categorization of emails and financial reports, improving efficiency by 94% on unstructured data.

LEADERSHIP & EXTRA-CURRICULAR ACTIVITIES

- Event Manager, Null Chapter Club – VITAP
- Member, Bulls and Bears Finance Club
- Volunteer, Intellect Fest 2023

CERTIFICATIONS

- Microsoft Certified: Azure Fundamentals (AZ-900)
- Neural Networks and Deep Learning – DeepLearning.AI (Coursera)
- Generative AI with Large Language Models – DeepLearning.AI (Coursera)