Vaibhav Deokar

vaibhav.deokar@outlook.com vaibhavdeokar.netlify.app linkedin/vaibhav-deokar /dev-ai-kar

Experience

Cognizant Technology Solutions – GenAI December 2023 – Present

Associate Data Scientist *Mumbai, India*

- Designed and built an autonomous AI Assistant for Google gTech Ads utilizing LLMs with grounding techniques on internal data sources like Plx to recommend Ads solutions to Ad agencies and patners; operated within Google's XWF development ecosystem using Bazel and Piper.
- Designed internal researcher tool with RAG, Neo4j Knowledge Graphs & Langchain.
- Built, deployed & secured GenAI Interview Bot (Azure GPT-4, Bedrock Claude 3.5) for HR; led red-teaming (jailbreaks, prompt injection) & MLOps (Kubernetes, CI/CD), validated with 1k employees.

Cognizant Technology Solutions - AI ML July 2022 – Dec 2023

Programmer Analyst *Mumbai, India*

- Led document retrieval system for a Johnson & Johnson client, leveraging RAG, GPT-3, and optimized Faiss vector search, boosting solution discovery rates by 61% across internal medical literature.
- Engineered an AI-driven system for Johnson & Johnson automating the categorization and semantic tagging of complex medical information requests using embeddings and t-SNE classification with Random Forest and K-Means (reducing response times by 57%) and accurately routing inquiries to subject matter experts with 86% precision.
- Leveraged advanced NLP techniques with tools like TensorFlow, PyTorch, Transformers, BERT, and T5 for complex data processing and contextual understanding across projects.
- Utilized models including GPT-2 and Salesforce CodeGen, optimizing inference speed with CUDA.
- Developed/optimized data processing pipelines (Python, Sklearn, Docker) & performed analysis (AWS S3/SageMaker, Pandas, NumPy) on 50+ datasets, enhancing efficiency.

Projects

Knowledge Graphs for RAG with Neo4j (/knowledge-graph-rag) | *Neo4j*, *LangChain*, *Python*

- Developed a Neo4j knowledge graph enhancing Retrieval Augmented Generation (RAG) accuracy; designed schema, extracted entities/relationships, implemented graph queries, and integrated with LangChain enabling precise hybrid retrieval from structured data.

Deploy ML model on REST API & Web App (/expert-enigma) | *Python*, *TensorFlow*, *Django REST API*, *React*

- Built and deployed a handwritten digit recognition CNN using TensorFlow, served via a Django REST API integrated with a React web application; explored deep learning image recognition using a dataset of 70,000 images.

Education

Syracuse University Expected Start: Fall 2025

Master of Science (M.S.) in Applied Human-Centered Artificial Intelligence

- **Relevant Coursework:** Generative AI Models, Deep Learning, NLP, Human-AI Interaction

Mumbai University Sept 2018 – Sept 2022

BS in Electronics Engineering

- **Relevant Coursework:** Programming, Database Management Systems, Computer Architecture, IoT

Skills

Programming & Data Science: Advanced proficiency in Python (including libraries such as NumPy, Pandas, SciPy, Scikit-learn), SQL, C / C++, Java, JavaScript and Bash scripting
AI/ML Research & Theory: CNNs, RNNs, LSTMs, Transformers, GANs, Variational Autoencoders, Diffusion Models, Reinforcement Learning, Optimization Algorithms, Benchmarking & Evaluation Metrics
Generative AI & LLMs: Expertise with LLMs (OpenAI GPT series, Claude, Gemini, Llama, BERT, T5), Diffusion Models, Multimodal Models, Prompt Engineering, Fine-tuning Strategies (LoRA, QLoRA, PEFT), Retrieval-Augmented Generation (RAG), Agentic Systems, Bias/Fairness/Safety Evaluation, Hugging Face and Langchain
MLOps & Cloud Infrastructure: Containerization and orchestration (Docker, Kubernetes), CI/CD pipelines (GitHub Actions), Git, AWS, Azure, GCP, Terraform, CloudFormation
Data Technologies & Web: Proficiency with SQL and NoSQL databases, Faiss, Pinecone, Neo4j, data visualization, Jupyter Notebooks, API design (gRPC, RESTful, GraphQL), React, Node.js, Django, Flask