SHUBHAM MIGLANI

miglanishubham25@gmail.com | GitHub | WebSite

WORK EXPERIENCE

Acentra Health, San Jose, CA | *Senior Machine Learning Engineer (Lead)* Jan 2024 - Present

- Developed enterprise-scale RAG infrastructure processing 100M+ documents annually, supporting 10,000+ monthly queries through AWS Bedrock Knowledge Bases, OpenSearch vector search, and S3-based ingestion pipelines.
- Shaped enterprise AI strategy and architecture, building shared agentic infrastructure using AWS Bedrock AgentCore, MCP Gateway, reusable agents, and unified APIs that enable AI capabilities across multiple product lines, reducing time-to-market for AI features by 50% through component reuse.
- Optimized retrieval cost and performance by tiering vector search - Amazon S3 Vectors for batch workloads (sub-second retrieval) and OpenSearch for real-time Q&A (p95 latency <50ms under concurrent load), reducing infrastructure costs by 40%.
- Applied advanced RAG techniques including query decomposition, hybrid retrieval, and cross-encoder reranking, improving answer accuracy by 15% (from 73% to 88%) in long-document question answering.
- Reduced manual document review time by 30% by building case summarization and Q&A systems using Bedrock LLMs to generate structured AI summaries from thousands of pages.
- Created an evaluation framework using judge LLMs and custom grounding metrics with manual feedback loops to continuously measure accuracy, factuality, and hallucination rates across production workloads.
- Led development of a Generative AI document automation tool with Azure OpenAI, FastAPI, and React, cutting letter drafting time by 50% and saving 11,000+ nursing hours (~$800K annually).
- Implemented Infrastructure as Code (IaC) using AWS CDK (Python) to automate provisioning and scaling, reducing deployment time by 60% and improving cloud resource efficiency.
- Mentored 5 engineers on GenAI systems and led AI enablement efforts through AI-DLC, introducing Codex and Kiro to improve team coding velocity and adoption.

Bridgera, Raleigh, NC | *Machine Learning Engineer* Apr 2021 – Dec 2023

- Architected and deployed 8+ conversational AI systems including a company-wide RAG chatbot using Azure OpenAI, LangChain, and FAISS, reducing average query response time by 60% and improving information access for call centers and business development teams.
- Designed and optimized a high-throughput AWS data pipeline processing 35M+ PDF pages annually using OCR and NLP, reducing processing time by 50% and costs by 40% through automated text extraction and intelligent bookmarking for case review.
- Led end-to-end Azure cloud migration of ML systems from on-premise infrastructure, establishing CI/CD pipelines with Azure DevOps and reducing deployment time by 70% through automated MLOps workflows.

MathWorks, Natick, MA May 2020 – Aug 2020

Intern in Engineering Development Group

- Developed code generation workflow for MATLAB to dynamically switch between OpenCV and Arm-Compute libraries, enabling optimized C++ code generation for image processing functions.

EDUCATION

NC State University, Raleigh, North Carolina Dec 2020
Master's in Electrical Engineering (specialization in Machine learning) GPA 4.00
Punjab Engineering College, Chandigarh, India May 2016
Bachelor's in Electrical Engineering GPA 8.85/10

TECHNICAL SKILLS

GenAI & Agentic Frameworks: LangChain, LangGraph, OpenAI Agents, Strands, LlamaIndex
MLOps & Observability: MLflow, Arize, Docker, CI/CD (Jenkins, GitHub Actions)
ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, Keras, Hugging Face Transformers
AWS: AgentCore, Bedrock, Textract, SageMaker, Lambda, Step Functions, API Gateway, S3, CDK, CloudWatch

SHUBHAM MIGLANI

miglanishubham25@gmail.com | GitHub | WebSite

Azure: OpenAI, Machine Learning Studio, Databricks, DevOps, Data Factory, Functions
Development: Python, FastAPI, React, JavaScript, Terraform
Data & Vector Stores: OpenSearch, Pinecone, Redis, FAISS, PostgreSQL, ChromaDB