Shubham Ubhe

AI/ML Engineer | Python Developer | Cloud Migration Specialist
Pune, Maharashtra (INDIA) | 8459296471 | ubheshubham.37@gmail.com | lhttps://www.linkedin.com/in/shubhamubhe/

PROFESSIONAL EXPERIENCE

Results-driven AI/ML Engineer with proven expertise in developing production-grade GenAI solutions. Specialized in RAG systems, LLMs, and vector databases with demonstrated success in building AI-powered tools that reduced data migration times by 40% and improved retrieval accuracy by 75%. Strong background in full-stack AI application development complemented by cloud migration expertise.

EDUCATION

Bachelor of Technology, **Vishwakarma Institute of Information Technology, India (2018 to 2022), GPA : 8.99/10**
HSC, **Engineering Science, Baburaoji Gholap College (2016 - 2018), Grade : A+**

CORE COMPETENCIES

- **GenAI & LLMs**: Retrieval Augmented Generation (RAG), Vector Embeddings, LLM Fine-tuning
- **Machine Learning**: Supervised/Unsupervised Learning, Model Deployment, MLOps
- **Data Engineering**: ETL Pipelines, Data Migration, Apache Spark, Big Data
- **Cloud Platforms**: GCP (BigQuery, Cloud Functions), AWS, Databricks
- **Languages & Tools**: Python, SQL, FastAPI, ReactJS, Docker, Kubernetes, GoLang

PROFESSIONAL EXPERIENCE

LTIMindtree

AI & Machine Learning Engineer | January 2022 - Present ***Specialized in AI research, development and implementation of GenAI solutions with progressively increasing responsibilities***

AI Development & Research (2023-Present)

- Led development of enterprise LLM platform processing 20M+ tokens daily with 99.5% reliability, reducing operational costs by $150K annually
- Implemented proprietary fine-tuning pipeline for Llama-3 and Mistral models, reducing hallucinations by 65% and improving domain-specific response accuracy by 78%
- Architected multi-agent RAG system with specialized agents for data retrieval, reasoning, and output generation, reducing complex query resolution time from hours to minutes
- Developed custom vector embedding solution optimizing for semantic search with 75% improved accuracy and 40% faster retrieval compared to keyword-based methods
- Created context-aware document processing pipeline using computer vision and NLP techniques, automating extraction of structured data from unstructured documents with 92% accuracy
- Implemented token optimization strategies reducing API costs by 35% while maintaining response quality
- Designed and deployed conversational AI systems with advanced memory management, achieving 87% user satisfaction on complex multi-turn interactions

Data Engineering & AI Integration (2022-2023)

- Spearheaded development of Canvas Eureka, reducing data migration time by 40% for Fortune 500 clients
- Led Generative AI implementation for Alcazar Accelerators, contributing to official Databricks partnership
- Built AI-powered data reconciliation tool validating 10TB+ of migrated data with 99.8% accuracy
- Created interactive Power BI dashboards reducing client decision-making time by 25%
- Implemented automated cloud migration pipelines for enterprise clients using Python and Spark
- Developed clustering and object sequencing solutions for optimized data migration
- Worked on US Airlines Case Study analyzing flight cancellation patterns using Python and Snowflake

KEY PROJECTS

Advanced RAG System for Enterprise Log Analysis | 2024

- Architected and implemented comprehensive RAG platform processing 500GB+ of log data daily with 30+ million embeddings
- Designed hybrid retrieval system combining BM25 and dense vector search, achieving 89% relevance score on complex technical queries
- Implemented custom prompt router intelligently selecting between 7 specialized prompts based on query characteristics
- Created adaptive chunking algorithm improving context window utilization by 42%
- Built real-time monitoring dashboard for model performance metrics, token usage, and cost optimization
- Technologies: LangChain, Hugging Face, LlamaIndex, Pinecone, Weaviate, Sentence Transformers, PyTorch, FastAPI, React

Multimodal AI for Document Intelligence | 2023-2024

- Developed end-to-end document processing system combining computer vision and NLP capabilities
- Implemented table extraction model achieving 94% accuracy on complex multi-page financial documents
- Created custom layout analysis pipeline for detecting and classifying document sections with 91% precision
- Built fine-tuned BERT model for entity extraction from specialized technical documents
- Designed knowledge graph integration for contextual entity resolution and data enhancement
- Technologies: PyTorch, Transformers, LayoutLM, Donut, BERT, Neo4j, FastAPI

Canvas Eureka: Automated Cloud Migration Tool | 2023

- Built automated migration tool that processed 50+ database migrations for enterprise clients
- Reduced migration errors by 85% through AI-powered validation
- Technologies: Python, Apache Spark, GCP, Databricks, Graph Algorithms

Data Lineage Tool | 2022-2023

- Developed tool that automatically mapped 10,000+ database object dependencies
- Reduced planning time for database migrations by 60%
- Technologies: Python, OrientDB, Graph Algorithms, SSIS Parser

ACHIEVEMENTS

- Won Hackathon for building RAG agent for log analysis (Cash prize: 75,000 INR)
- Winner in VishwaCTF Challenge (Cash prize: 5,000 INR)
- Solvathon Finalist in LTIMindtree (Cash prize: 1,600 INR)

CERTIFICATION

- PCAP: Certified Associate in Python Programming (Python Institute)
- DeepLearning.AI: Building and Deploying LLM Applications
- Hugging Face: Natural Language Processing Specialist
- Google Cloud: Generative AI Fundamentals
- AWS: Machine Learning Specialty (In Progress)
- Data Warehousing for Partners: Enable Google Cloud Customers
- Programming for Everybody (Getting Started with Python)
- Introduction to Image Generation
- Advanced Operations Using Amazon Redshift
- Python Data Structures
- GenAI Foundation Course
- NVIDIA: Building Transformer-Based Natural Language Processing Applications