Sumaiya Ibrahim Shaikh

Los Angeles, CA | shaikhsumaiyaibrahim@gmail.com | +1-562-208-8511 | LinkedIn | GitHub

Experience

California State University, Dominguez Hills Jan 2025 – May 2025
Research Assistant, Carson, CA

- Researched and implemented **low-resource Neural Machine Translation (NMT)** models for clinical text (25K+ bilingual Nepali–English sentences), improving BLEU by 15%.
- Fine-tuned **mBART & NLLB-200** using PyTorch with FP16 mixed-precision on A100 GPUs, reducing training time by 30% while maintaining accuracy.
- Designed and executed **error profiling & ethical risk analysis**, integrating **human-in-the-loop workflows** for safer deployment in clinical translation.
- Established a **benchmarking framework** with BLEU, CHRF++, BERTScore, COMET, and Perplexity, improving reproducibility and enabling cross-model comparisons.
- Co-authored a **peer-reviewed research paper (RANLP 2025)**, presenting findings on domain adaptation in low-resource NMT.

Accenture Sep 2022 – Aug 2024
Software Engineer, Pune, India

- Designed and optimized **ETL pipelines in Informatica**, cutting data load time by 20% and increasing reporting performance by 25%.
- Automated large-scale ingestion workflows using **AWS S3 + Kafka + PySpark**, enabling real-time processing of 20M+ daily records and reducing manual interventions by 40%.
- Implemented **CI/CD pipelines** (GitHub, Docker, Concourse, Rancher), achieving zero-downtime deployments and reducing release cycles from weekly to daily.
- Enhanced **QA automation** by integrating Tosca regression suites into Jenkins, cutting ETL job failures by 50%.
- Delivered a **Databricks-based Data Lake** with Delta Tables and Spark SQL, boosting query performance by 30% and enabling downstream AI/ML pipelines.
- Migrated **Python 2.7 to 3.x** across production systems, resolving Informatica–Teradata connectivity issues and strengthening governance with Trillium-based quality checks.

Technical Skills

Programming & Scripting: Python, Java, JavaScript, UNIX, SQL, PL/SQL, MongoDB, SQL Server, DynamoDB
Machine Learning & AI: NLP, Neural Networks, Deep Learning, Reinforcement Learning, LLM Integration (OpenAI API, Hugging Face), LangChain, LangGraph, TensorFlow, PyTorch, Scikit-learn, Pandas, NumPy, Ethical AI
Data Engineering & Big Data: ETL Development (Informatica, PySpark, Databricks), Apache Spark, Hadoop, Delta Lake, Snowflake, Kafka, Data Lakes, Data Warehousing
Cloud & DevOps: AWS (S3, Databricks, API Gateway), GCP, Oracle Cloud, Docker, Kubernetes, Terraform, CI/CD (GitHub Actions, Jenkins, Concourse, Rancher)
Web & API Development: Django REST Framework, Spring Boot, Express.js, React.js, Tailwind CSS, OAuth2, GraphQL, RESTful APIs, SOAP APIs, Postman, Swagger, Microservices
Software Engineering Practices: Agile (Scrum, Kanban), Software Development Life Cycle (SDLC), Software Quality Assurance (SQA), Test-Driven Development (TDD), Behavior-Driven Development (BDD)

Projects

Hyper-Personalized Financial Product Engine (Work in Progress) 2025
Python, AWS Bedrock, LangChain, Streamlit, OpenAI GPT-4o, DynamoDB

- Developing an **AI-powered recommendation system** that personalizes financial products (loans, credit cards, savings plans) based on user profiles and real-time data.
- Integrating **LLMs via LangChain and AWS Bedrock** to generate natural-language explanations for each recommendation, enhancing user trust and transparency.
- Building a **modular backend** using Python APIs and DynamoDB for scalable data handling, coupled with a Streamlit-based interactive dashboard.
- Designing a **retrieval-augmented pipeline** that learns from feedback, enabling adaptive and explainable decision-making in financial contexts.

AI Email & Task Sorter 2025
Python, OpenAI GPT-4o-mini API, Gmail IMAP, Todoist API

- Built an **automation pipeline** integrating Gmail IMAP with GPT-4o-mini, automatically classifying, summarizing, and routing 1K+ emails into Todoist tasks.
- Implemented **secure multithreaded IMAP ingestion with SSL**, reducing processing latency by 40% while ensuring privacy.
- Combined **LLM-based classification** with fallback rule-based logic, achieving 98% task creation accuracy for workflow automation.

Digital Currency Price Prediction 2023

Python, LSTM, Scikit-learn — Published in IJRSET

- Engineered a **time-series forecasting pipeline** using LSTM neural networks with real-time exchange API integration, outperforming regression baselines.
- Achieved **95% prediction accuracy ($R^2 > 0.99$)** on validation sets, enabling improved investment insights.
- Published research findings in **IJRSET 2023**, highlighting deep learning applications in financial forecasting.

Certifications

- Oracle Cloud Infrastructure 2022 Foundations Associate

Education

California State University, Dominguez Hills (CSUDH) Expected May 2026

Master of Science in Computer Science; GPA: 3.86/4.00

Relevant Coursework: Machine Learning, Advanced Programming Languages, Cloud Computing, Software Quality Assurance,NLP.