

Mohammed Shaik

214-430-8536 | Mohammed27shaik@gmail.com | [Mohammed-Shaik.com](https://www.linkedin.com/in/mohammed-shaik-24268b425/) | [linkedin.com/in/mohammed-shaik-24268b425/](https://www.linkedin.com/in/mohammed-shaik-24268b425/) | github.com/arsalan5shaik

EDUCATION

The University of Texas at Dallas — Richardson, TX

Expected Dec 2026

Bachelor of Science in Computer Science | GPA: 3.8

Relevant Coursework: Machine Learning, Artificial Intelligence, Computer Vision, Advanced Algorithm Design and Analysis, Database Systems, Linear Algebra, Probability & Statistics

TECHNICAL SKILLS

Languages: Python, Java, SQL, TypeScript, C++, R, Go

Machine Learning: PyTorch, TensorFlow, Keras, scikit-learn, NumPy, Pandas, OpenCV, CNNs, transformers, transfer learning, fine-tuning, computer vision, NLP

LLMs & Generative AI: Whisper, Hugging Face, LangChain, RAG, hybrid search, reranking, agents & tool calling, structured outputs, embeddings

Evaluation & Optimization: Golden-set and model-graded evals, regression testing in CI, semantic caching, model routing, token & latency tracing

Vector & Databases: pgvector, FAISS, PostgreSQL, MongoDB, Redis

MLOps & Cloud: AWS (SageMaker, S3, Lambda, Redshift), Docker, Kubernetes, MLflow, Airflow, FastAPI, CI/CD, Linux

Data Engineering: Kafka, Apache Beam, ETL pipelines, streaming data, feature pipelines, data validation

WORK EXPERIENCE

Staples, Framingham, MA

Jun 2026 — Aug 2026

AI Engineer Intern

- Built and deployed an LLM service in Python and FastAPI that handles internal operations requests end to end: it reads the request, calls internal APIs for the data it needs, and completes the action, replacing manual triage that cost 15 hours a week.
- Designed the tool-calling layer that lets the model choose among 8 internal APIs, validating arguments against a schema, retrying invalid calls, and falling back to a deterministic path, holding success above 99% across ~8,000 requests a day.
- Added retrieval over internal policy documents with hybrid keyword and vector search plus reranking, raising accuracy from 78% to 94% on a 500-question labeled set and cutting hallucinated values by 60%.
- Built an evaluation harness with a labeled golden set and model-graded scoring that runs in CI on every prompt or model change, blocking merges that regress accuracy, latency, or cost.
- Cut inference cost by 35% by routing simple requests to a smaller model and caching repeated lookups, and added per-request token and latency tracing to find the expensive paths; deployed with Docker on Kubernetes.

Intuit, Plano, TX

Feb 2025 — Aug 2025

Software Engineer Intern, Data Platform

- Built Python components of a streaming pipeline handling about 40 million events a day from Kafka, turning raw events into user-session features for downstream machine learning models.
- Wrote features into a shared store so training and serving read the same definitions across 10+ event types, and built a monitoring tool that flags when the pipeline falls behind and recommends how much to scale up, cutting peak-hour delays by 30%.
- Moved 15+ jobs to AWS with S3 and Redshift, scheduled them with Airflow, and wrote validation scripts confirming no records were lost, improving query speed on training data by 25%.

PROJECTS

ScreenGuide — Multimodal AI Desktop Assistant

- Built and deployed a multimodal assistant for Windows and macOS that answers spoken questions about the user's screen, pairing GPT-4o vision with Whisper speech-to-text; grew to 9,000+ users and \$600+ in revenue over eight months.
- Built the vision and speech pipeline in Python: screen capture, image resizing and tiling, streamed responses from GPT-4o vision, and Whisper transcription, with a median 1.2s to the first word.
- Cut inference cost by 45% with a semantic cache that reuses answers to similar questions by embedding similarity, plus an intent classifier routing 40% of requests to a smaller model.

Global Intelligence Dashboard — Live News & Risk Monitoring

- Built and deployed a live dashboard that pulls in 200+ news, market, and disaster feeds, clusters related stories by embedding similarity, and produces a schema-validated summary and risk score per cluster across 30+ countries.
- Wrote one interface that works with either a hosted LLM API or a local Ollama model, so the same pipeline runs on a laptop or in production.
- Cut LLM costs by 60% by deduplicating near-identical stories before summarizing and caching results in Redis; containerized with Docker.

Sign Language Recognition — Video Classification

- Trained a video classification model in PyTorch on the WLASL dataset (21,000 videos, 2,000 signs), combining a pretrained I3D backbone for motion features with a six-layer Transformer encoder for sequence modeling.
- Improved top-1 accuracy from 65.9% to 74.5% over the I3D baseline, measuring with top-1 and top-5 accuracy, precision, recall, and confusion matrices.
- Fixed class imbalance with weighted sampling and a weighted loss, and benchmarked a Vision Transformer backbone that overfit on smaller subsets.

FillRight — LLM-Assisted Job Application Tool

- Built and deployed a tool that uses an LLM to write open-ended job application answers from a saved profile, returning structured output that maps onto form fields; cut a 12-minute application to under 90 seconds across 400+ applications.
- Routed most requests to a smaller, cheaper model and fell back to a second LLM provider when a response failed schema validation.

CERTIFICATIONS

PCEP — Python Programmer | Microsoft Azure Fundamentals (AZ-900) | Azure Data Fundamentals (DP-900) | Google Cloud Professional Data Engineer - In progress