

Ali Rahimian

+1 619-904-5635 | akhalegh@charlotte.edu | linkedin.com/in/alikrahimian | github.com/ak811 | [Scholar](#)

Software & Machine Learning Engineer | Charlotte, NC

SUMMARY

Software engineer with 8+ years of experience across ML research, production software, and data engineering in Python, Java, Kotlin, and SQL. At the Charlotte Machine Learning Lab, led Fibottention as first author, an $O(N \log N)$ attention module with $2.45\times$ lower latency and 70% less GPU memory, and co-developed TruthLens (ICML 2025). As founder and sole engineer of IRSoft, built and operated mobile apps serving 3.5M+ users with a 4.7/5 rating; built Spark and AWS pipelines over 3M+ records.

TECHNICAL SKILLS

Languages: Python, Java, Kotlin, SQL, C/C++, CUDA, JavaScript, Dart, PHP, R, Bash
Machine Learning & LLMs: PyTorch, Hugging Face Transformers, Scikit-learn, Spark MLlib, NumPy, Pandas, Vision Transformers, Vision-Language Models, LLM-as-a-Judge Evaluation, Recommender Systems
ML Systems & HPC: CUDA, OpenMP, PyTorch DDP, SLURM, Mixed Precision, Int8 Quantization, Batch Inference, Evaluation Pipelines, Provenance Logging, Latency/Throughput/Memory Profiling
Data Engineering & Cloud: PySpark, Spark SQL, Structured Streaming, Hadoop MapReduce, Kafka, Trino (Amazon Athena), Parquet, AWS (S3, Lambda, Glue, Athena, EC2, IAM), MySQL, PostgreSQL, MongoDB
Backend & Mobile: REST APIs, Flask, Node.js (Fastify), React, JWT Authentication, Android (Kotlin, Java, Jetpack Compose), Kotlin Coroutines and Flow, Retrofit/OkHttp, Firebase, Flutter (Dart)
DevOps, Testing & Security: Docker, Docker Compose, CI/CD (GitHub Actions), Linux, Pytest, Hypothesis (property-based testing), JUnit, AES Encryption, RSA Receipt Verification, R8/ProGuard Hardening

EXPERIENCE

Machine Learning Engineer (Research Assistant)

Jan. 2024 – Dec. 2025

Charlotte Machine Learning Lab, UNC Charlotte

Charlotte, NC

- **Batch Inference Pipeline (TruthLens, ICML 2025):** Engineered a resumable probe→aggregate→judge→evaluate pipeline that let GPU inference and API-based LLM scoring run on separate machines, with append-on-write JSONL checkpoints, deterministic seeding, and provenance logging.
- **Explainable AI (TruthLens, ICML 2025):** Co-developed a training-free, explainable detection system in which vision-language models answered nine probes and an LLM judge returned a verdict with a justification.
- **Reliability & Testing:** Made the pipeline testable without GPUs or live API calls: wrapped four vision-language models as pluggable backends with fault-tolerant LLM output parsing, and covered prompts, aggregation, parsing, metrics, and resumption with a Pytest suite, a local mock API server, and a CPU-only smoke test.
- **Model Evaluation:** Benchmarked TruthLens against CnNDetection and DIRE: implemented balanced accuracy, AUC, precision, recall, and F1 in NumPy, validated them against scikit-learn, and ran ablations isolating the contribution of prompting, LLM reasoning, and model backbones.
- **Performance Engineering (Fibottention):** Led development of a drop-in PyTorch attention module for vision transformers cutting cost from $O(N^2)$ to $O(N \log N)$; profiled latency, throughput, and peak GPU memory against dense attention, measuring $2.45\times$ lower latency, $2.49\times$ higher throughput, and 70% less memory.
- **Distributed GPU Training:** Ran 8-GPU training with PyTorch DDP, `torchrun`, SLURM, mixed precision, and checkpoint/resume; set up and migrated the lab's training environment to a university HPC cluster.
- **Reusable ML Tooling:** Built a config-driven benchmarking suite (registry/factory design, YAML configs) that swapped 10 attention mechanisms without code changes; ran 100+ experiments with multi-seed variance analysis.

Software Engineer & Founder

Mar. 2020 – Dec. 2023

IRSoft

Mashhad, Iran

- **Production Ownership:** As the sole engineer, designed, built, and operated consumer apps serving 3.5M+ users with a 4.7/5 rating across 120K+ reviews, owning architecture, backend, database, and releases.
- **Clients & REST Backend:** Kept core features working offline from first launch by building native Java and Kotlin clients on a PHP REST API and MySQL backend with Retrofit/OkHttp networking, offline-first SQLite persistence, and minimum-version enforcement to retire outdated clients.
- **Security & Fraud Prevention:** Verified in-app purchase receipts with RSA signatures to reject spoofed transactions, encrypted user data with AES, and hardened builds with R8/ProGuard.
- **Monitoring & Root Cause Analysis:** Tracked post-release stability with Crashlytics, handled network and page-load failures with graceful recovery, and built on-device usage analytics that drove roadmap decisions.

Software Engineer

Jan. 2019 – Feb. 2020

DIGM

Yazd, Iran

- **Leadership & Agile:** Led Java/Kotlin applications through Scrum releases; defined code-review standards and a Git branching strategy, integrated JUnit tests into the Gradle build, and mentored junior developers and interns.
- **Real-Time Rendering & Performance:** Architected a reusable Java/OpenGL ES rendering framework sustaining 60fps with hundreds of on-screen objects; removed frame-time spikes and garbage-collection pauses through CPU, GPU, and memory profiling and object pooling.

Data Engineer

Jun. 2019 – Nov. 2019

Yazd University

Yazd, Iran

- **ETL & Data Quality:** Replaced manual spreadsheet workflows with reproducible SQL and R pipelines and added automated validation checks for grade ranges, credit totals, and enrollment integrity.
- **Statistical Analysis:** Ran cohort analyses with t-tests, chi-square tests, and logistic regression to evaluate prerequisite-course policy, flagged at-risk students from indicators such as repeated failures and declining GPA, and presented recommendations to non-technical administrators.

Software Developer & Co-founder

Mar. 2018 – Feb. 2019

Pishgaman Accelerator

Yazd, Iran

- Took a startup product from concept to MVP: a Java client and a RESTful PHP backend on a normalized MySQL schema with token-based authentication, query optimization, pagination, and response caching.

SELECTED PROJECTS

Sparxiv: Recommender System over 3M+ Records | *PySpark, Spark ML, Parquet, TF-IDF, SciPy CSR*

- Built an end-to-end recommender over the full arXiv metadata snapshot: JSONL to year-partitioned Parquet ETL, a Spark ML TF-IDF text-feature pipeline, and SciPy sparse similarity search in place of Spark cross joins.

Ase: Multilingual Search Engine | *Java, BM25, Positional Index, External Merge Sort, Docker, REST API*

- Built a self-hosted search engine in Java serving a web UI and JSON API, indexing 200,000 documents in 28s under a 384MB heap and answering typical queries in 1–12ms, with BM25 ranking and atomic hot index reload.

Finalytics: Financial Transaction Analytics | *Python, Pandas, SciPy, Vectorized ETL, Anomaly Detection*

- Built a validated ETL that unified transaction exports from multiple financial institutions into a canonical schema, with rule-based categorization, merchant entity resolution, transfer deduplication, and z-score anomaly detection across 17 reports.

Distributed Data Pipelines | *AWS (S3, Lambda, Glue, Athena), Spark Streaming, Spark MLlib, Trino SQL*

- Architected an event-driven S3→Lambda→Glue→Athena (Trino) ETL with least-privilege IAM and a Flask/boto3 dashboard on EC2; wrote Athena SQL with CTEs and window functions.
- Built a real-time PySpark Structured Streaming pipeline with watermarked event-time sliding windows that scored every incoming event with a persisted Spark MLlib regression model.

EDUCATION

University of North Carolina at Charlotte

Charlotte, NC

Master of Science in Computer Science; GPA: 3.9/4.0

Jan. 2024 – Dec. 2025

- Direct admit to the Ph.D. track; completed doctoral coursework and NSF-funded research; graduated with the M.S.
- Thesis on multi-head sparse attention for efficient Transformers; advised by Dr. Christian Kümmeler.

Yazd University

Yazd, Iran

Bachelor of Science in Computer Software Engineering; GPA: 3.7/4.0

Aug. 2017 – May 2022

- Ranked 1st for three consecutive years; two-time Shahrard Merit Scholar; TA for three courses and 130+ students.

PUBLICATIONS

TruthLens: Training-Free Data Verification for Deepfake Images via VQA-style Probing. *ICML 2025*.

Fibottention: Inceptive Visual Representation Learning with Diverse Attention Across Heads. *arXiv 2026*, first author.

HONORS & AWARDS

Competitions: Honorable Mention, 17th ACM ICPC Asia Tehran Regional (731 teams, 94 universities); 3rd Place, Sharif ICT Fintech Hackathon (65 teams; ranked 1st on the top-weighted challenge by processing 5.7M+ transactions in under 10s); 4th Place, Sharif Obfuscated C Code Contest (200+ participants); 10th Place, JavaCup National Finals; 1st Place, Yazd University Chess Championship; 2nd Place, Razavi Khorasan Provincial Chess Championship (U-14).