kinzayousaf.001@gmail.com
linkedin.com/in/kinza-y/
github.com/Elasticminds

Kinza Yousaf

647-766-4493
Toronto, ON

OBJECTIVE

Seasoned AI Software Engineer with 5+ years of industry experience and a strong business focus. I've helped early-stage startups go from zero to production owning ML infrastructure, deploying real-time personalization engines, and launching AI products now used by thousands. From building MVPs to scaling backend systems and GenAI pipelines, I help turn prototypes into revenue-generating products.

SKILLS

- **Cloud & DevOps:** AWS, Azure, GCP, Docker, Kubernetes, Terraform, Helm, Jenkins, GitHub Actions, Redis, Linux
- **Programming Languages:** Python, Java, Scala, JavaScript, SQL (MySQL, PostgreSQL, OracleDB), MongoDB
- **Data & ML Ecosystem:** Spark, Hadoop, Kafka, NumPy, Pandas, PyTorch, TensorFlow, OpenCV, WandB, R
- **AI/ML Algorithms:** LangChain, LlamaIndex, ElasticSearch, Zero-shot/Few-shot Learning, Gen AI, BERT, GPT, Gradient Boosting, Gradient Descent, Transformers, Flask, Seaborn, LaMDA, LLaMa, Huggingface and AutoGPT/AgentGPT

EXPERIENCE

AI Engineer, Hoja AI — Jan 2025 - Present

- Collaborated directly with the CTO from product conception through launch, driving architectural decisions that enabled 2× faster feature iteration and ensured scalability for 10k+ monthly active users.
- Architected and deployed Hoja's adaptive-learning AI pipeline, developing data ingestion, model training orchestration and inference-at-scale on GKE; resulting in a 34% boost in content personalization accuracy.
- Designed Hoja's student-agent interaction pipeline, handling 5k+ concurrent users with sub-150 ms response times to dynamically adjust difficulty and learning paths, boosting average session duration by 27% and retention by 21%.
- Showcased Hoja's human-centered AI at ITC Innovation Fair; helped earn Top-15 Generative AI Startup ranking on F6S.
- Build a lightweight experimentation SDK in Kotlin that ties into your feature-flagging service LaunchDarkly, so product can spin up "personalized learning" vs. "control" experiences.

Machine Learning Engineer, cPacket — Sept 2023 - Jan 2025

- Built an LLM agent using a fine-tuned Llama 3.18B Model with parameter-efficient QLoRA and PEFT optimizations, designed to analyze network flow data and detect DDoS attack patterns by interpreting packet loss in network traffic.
- Containerized training and inference with Docker; wired experiments and model versions into MLflow.
- Transformed cVu-NG into the industry's first self-driving packet broker, deploying a fully automated ML backbone that processes every packet in real time, surfaces critical anomalies and maintains sub-50 ms tail-latency under 200 Gbps loads.
- Trained and deployed a multi-class Deep Neural Network using TensorFlow to automatically tag each packet flow identifying key protocols, encrypted vs. cleartext traffic and potential security threats with 98% accuracy, enabling security tools downstream to focus on the right data.
- Took raw, line-rate packet streams from every port of the cVu-NG ASIC/FPGA nodes into a Kafka pipeline, built feature transformers like microburst signatures and inter-packet timing histograms in Spark Streaming, and exposed real-time anomaly scores via a gRPC inference service running on Kubernetes.
- Recognized by the VP of Engineering for delivering an AI-driven traffic distribution solution that cut packet loss by 80% under peak loads.

AI Software Engineer, Next Play — Dec 2022 - May 2023

- Led a team of 5 interns to develop and deploy full-stack features including secure authentication flows and a real-time notification center for a new gaming platform using the PERN stack (PostgreSQL, Express, React, Node) and AWS RDS.
- Architected and built the Next Play Android app in Kotlin, implementing live pitch-prediction UIs, real-time leaderboards, delivering sub-100 ms game updates and driving a 31% increase in daily active users.
- Streamlined releases through Git-based workflows, performed code reviews and coordinated sprint deliverables.
- Designed and implemented Node.js microservices for prediction scoring, points accounting, and friend-matchmaking, supporting over 5,000 concurrent users with 99.8% uptime and reducing server response times by 60%.
- Integrated real-time analytics and A/B testing to optimize game mechanics and UI flows, boosting first-week retention.

Data Scientist, Afiniti Jan 2020 - Nov 2022

- Led the AI Production team in designing and deploying end-to-end ML pipelines, from feature engineering to model serving, for enterprise clients across Europe.
- Delivered data-driven solutions that achieved 3x revenue lift and 5x ROI, leveraging real-time statistical models and XGBoost for clients like NOS Portugal and Telefonica Spain.
- Deployed models to production with robust MLOps practices, including drift detection, high availability and continuous performance monitoring.
- Designed and developed a scalable healthcare ecosystem, integrating AI-based solutions while ensuring compliance with HIPAA, GDPR and data governance standards.

EDUCATION

Carnegie Mellon University (CMU), Master of Information Systems Management Aug 2023 - Dec 2024

- **Relevant Courses:** Distributed Systems, Cloud Computing, Java Application Programming, Generative AI, Machine Learning with Large Datasets, DevOps and Continuous Integration
- Cumulative GPA: 3.94

Lahore University of Management Sciences, Bachelor of Science in Computer Science Aug 2016 - Dec 2020

- **Relevant Courses:** Calculas, Linera algebra, Computer Networks, Computer Vision, Applied Probability, Advanced Programming and Data Mining

PROJECTS

MedSync AI – Intelligent Healthcare Workflow Automation Suite

An AI-powered healthcare automation suite that streamlines clinical documentation, enhances workflow efficiency, and improves decision-making.

- Automated Clinical Documentation – Generates SOAP notes, discharge summaries using NLP and speech recognition, reducing physician workload.
- Medical Transcription – Converts doctor-patient conversations into structured EHR entries, ensuring compliance.
- AI-Driven Prior Authorization – Automates insurance approvals by validating medical necessity, cutting processing time.
- Medical Chatbot – Provides real-time, ontology-based responses to medical queries, assisting healthcare professionals.

Key Contribution : Reduced physician admin workload by 37%, accelerated insurance approvals by 29%, and improved documentation accuracy. Solved manual data entry inefficiencies and billing errors in healthcare.

Prompt-to-Prompt Image Editing using Generative AI Foundation Models

- Implemented a latent diffusion pipeline using PyTorch and HuggingFace Diffusers, optimizing cross-attention mechanisms for image generation.
- Engineered token replacement and attention map manipulation to enable text-driven image editing.

Prompt-to-Prompt Image Editing using Generative AI Foundation Models

- Designed and deployed a highly available cloud-native microservices system on AWS using EKS (Kubernetes), RDS, S3, and multi-layer load balancing, provisioned via Terraform and Helm.
- Implemented multi-phase ETL with Apache Spark to process and migrate 1TB+ of Twitter data from Azure Databricks to AWS, tuning infrastructure to optimize throughput to cost ratio.
- Enabled secure inter-service communication in a resilient, multi-protocol (REST/gRPC) architecture with automated CI/CD pipelines using GitHub Actions.

Cloud-Native Microservices System for Blockchain, QR Code, and Twitter Recommendations

- Designed and deployed a highly available cloud-native microservices system on AWS using EKS (Kubernetes), RDS, S3, and multi-layer load balancing, provisioned via Terraform and Helm.
- Implemented multi-phase ETL with Apache Spark to process and migrate 1TB+ of Twitter data from Azure Databricks to AWS, tuning infrastructure to optimize throughput to cost ratio.
- Enabled secure inter-service communication in a resilient, multi-protocol (REST/gRPC) architecture.