Rafay Khan

Principal AI Engineer

rafayak.github.io |+92 346 540 2652 | rafayak16@gmail.com | GitHub | LinkedIn

Principal AI/ML Engineer with 8+ years of experience building production AI systems across applied LLMs, computer vision, recommendation systems, and edge inference. Led end-to-end model development, evaluation, deployment, and monitoring for real-world AI applications, including multilingual customer-support assistants, high-throughput inference stacks, real-time video intelligence, and large-scale edge deployments.

PROFESSIONAL EXPERIENCE

VisionX – *AI development and consulting firm building custom ML solutions*

Principal AI Engineer | May 2024 – Present *Dubai, UAE / Islamabad, Pakistan*

- Serving as technical lead across multiple engineering workstreams building an **LLM-powered customer support assistant** for **Iraq's largest payments processor**; supports **Arabic, Kurdish, and English** across **chat + real-time voice**, handling **50,000+ unique conversations/day** for a platform serving **30M+ end customers**.
- Delivered **~5s end-to-end response times** with autoscaling up to **2,500+ requests/sec** on **AWS ECS**, infrastructure managed via **Terraform**.
- Built production inference stack using **FastAPI, OpenAI and custom models via SGLang LLM Inference Server,** supporting **multimodal inputs** (voice, images, and text) and real-time call flows.
- Built runtime **guardrails** and fallback logic for a multilingual agentic support assistant, controlling low-confidence responses, tool/API failures, unsafe flows, and human escalation across Arabic, Kurdish, and English conversations.
- Developed offline **evaluation suites** and production monitoring to test prompt/model changes, track resolution quality, analyze escalation patterns, and catch regressions before deployment.
- Instrumented the entire application stack with **OpenTelemetry** traces, metrics, and logs to debug chat and real-time voice flows across model calls, backend services, inference endpoints, guardrail decisions, and human handoff paths.
- Defined **OpenAPI contracts** with backend teams and enabled **human-in-the-loop** escalation flows for low-confidence, high-risk, or unresolved customer interactions.

Lightly.ai – *YC-backed data curation platform; Lightly Edge selects high-value data on-device*

Senior Machine Learning Engineer (Consultant) | January 2024 – April 2024 *Zürich, Switzerland*

- Contributed to **LightlyEdge**, an on-device **smart data selection SDK** that identifies high-signal samples in real time to reduce **data transfer and storage** overhead in edge ML pipelines.
- Implemented edge inference components in **Rust** using **ONNX Runtime** to run model-driven selection logic efficiently under strict runtime constraints (Nvidia Jetson Xavier).
- Designed and integrated **model-serving primitives** (pre/post-processing, batching, interface boundaries) to support reliable inference inside a production SDK workflow.
- Contributed **to model optimization** through **quantization and pruning** workflows, helping reduce edge inference overhead while preserving predictable selection behavior under resource-constrained deployments.

Veeve.io – *Smart shopping-cart retail tech using computer vision + sensor fusion*

Senior Machine Learning Engineer – Team Lead | January 2020 – November 2023 *Seattle, USA*

- Led ML development for a **real-time in-cart video intelligence system** spanning action detection and product tracking, running at **20 FPS** across production edge deployments.
- Owned **model development, evaluation, and annotation-team workflows** for action detection and product tracking, balancing **precision/recall, latency, and edge constraints** across noisy retail environments.
- Scaled inference across **1,500+ smart carts running simultaneously** across **USA, Europe, and the Middle East**, delivering stable real-time decisioning under noisy retail conditions.
- Productionized model serving on **GCP + Kubernetes**, scaling custom CV models via **NVIDIA Triton Inference Server** for high-throughput, low-latency inference.
- Built an end-to-end **ML deployment pipeline** with **GitOps-based releases** and controlled model rollouts, enabling safe promotion of model versions and faster production iteration.
- Implemented experiment tracking and reproducibility workflows with **Weights & Biases** (training runs + dataset + artifact/version tracking), supporting faster debugging and reliable model lifecycle management.
- Supported deployments with major US grocery retailers including **Albertsons** and **Kroger**, contributing to large-scale operational rollout readiness.
- Work contributed to product viability and investor outcomes, including a publicly reported **$6.7M funding round**.

▶ Watch 2-minute early prototype demo of real-time in-cart activity detection

(In this video, the model identifies seven fine-grained actions including insert, remove, rummage, and theft-like behavior. Visualized via a step-function and live confidence bar chart. Since the demo, I expanded class coverage, especially for theft detection, and optimized accuracy for both edge and cloud deployment.)

Presidential Initiative for Artificial Intelligence and Computing (PIAIC)

Assistant instructor *(part-time)* ***| September 2019 – September 2020*** *Islamabad, Pakistan*

- Selected as an **early PIAIC cohort admit (top ~1.7%)**, later serving as **Assistant Instructor** supporting learners, designing assignments and helping with technical reviews.

SPECIAL PROJECTS

Skinsight.me – *Personalized AI-driven skincare product recommendation platform*

Co-Founder *Dubai, UAE*

- Built an AI skincare recommendation product end-to-end, spanning **frontend in Next.js**, **backend in FastAPI**, **Temporal-managed durable data ingestion pipelines** for continuous catalog growth, and an **agentic product discovery assistant** that recommends skincare products through conversational flows; all **scaled on AWS ECS**.
- Trained a **neural collaborative filtering system** on **60K+ products** and **2M+ user data points** to generate personalized recommendations and high-intent alternatives.
- Designed architecture for scalable iteration: modular pipelines, reproducible training runs, and clear interfaces between data acquisition, modeling, and product delivery.

RESEARCH EXPERIENCE

Mila

Research Fellow *(part-time)* ***| April'22 – June'24*** *Quebec, Canada (Remote)*

Research Fellow as part of the Fatima Fellowship. The primary investigator on the project ***"More than more data? Comparing training with data augmentation vs IID data"***. The project aims to dissect the effect augmented samples have on representational & adversarial robustness, and model convergence in various domains.

Supervisor: Alex Hernandez-Garcia

- Designed controlled experiments comparing augmentation regimes against IID training to analyze effects on robustness, convergence, and learned representations.
- Built a modular Python experiment workbench (https://github.com/alexhernandezgarcia/data-augmentation) for reproducible training, evaluation, and ablation studies across progressively harder synthetic tasks.
- Developed custom augmentation methods and toy datasets to isolate how specific transformations affect model behavior and parameter learning.

TUKL-NUST R&D Lab

Research Assistant | Sep'16 – June'17 *Islamabad, Pakistan*

Vision research team lead for the project ***"Fish Biodiversity Estimation by Low-Cost Non-Destructive Video Based Sampling (FIBEVID)."*** This project was in collaboration with Pakistan Fisheries Department, University of Kaiserslautern (TUKL) to help create a non-destructive fish bio-diversity estimation tool for Pakistan's National fish – the Mahasher. Project went on to receive funding from Hochschule RheinMain and DAAD (German Academic Exchange Service).

Supervisors: Dr. Faisal Shafiat, Dr. Imran Malik
(https://www.hs-rm.de/de/fachbereiche/design-informatik-medien/forschungsprofil/fibevid/)

- Created and published the first ever Mahasher fish dataset. The final dataset has 100,000 labeled images of Mahasher fish in its natural river habitat.

Research Assistant | June'16 – Sep'16 *(part-time)*

Summer intern at the TULK-NUST R&D lab. Worked in collaboration with researchers at University of Kaiserslautern detect forged bank cheques under the project "***Forgery detection and segmentation of handwritten text through Hyperspectral Image Analysis***."

Supervisor: Dr. Imran Malik

- Performed statistical analysis on hyperspectral image data to separate out inks of different varieties.

EDUCATION

National University of Sciences and Technology (NUST) *Islamabad, Pakistan*

Bachelor of Science in Computer Science (BSCS) ***2013 – 2017***

Rector's High Achiever's List 2017 Final Year Project - Fish Biodiversity Estimation by Low-Cost Non-Destructive Video Based Sampling, FIBEVID

SKILLS

Programming Languages: Python, C/C++, Rust, Java, Javascript/Typescript, Matlab
Deep Learning / LLM Frameworks: PyTorch, TensorFlow, Keras, LangChain / LangGraph, PydanticAI
Web Frameworks: Laravel, ReactJS, D3.js, Next.js, Astro
Cloud / DevOps / MLOps: AWS, GCP, Azure, Terraform, GitHub Actions, Docker, Kubernetes, SGLang, NVIDIA Triton, ONNX Runtime, Weights & Biases
Languages: Urdu (native); English (fluent), Spanish(basic), German (basic).

SELECT HONORS & AWARDS

- **Fatima Fellowship, 2022**. Selected in top 53 fellows from an applicant pool of over 700 from 68 different countries
- **Rector's High Achievers List, 2017** final year project – FIBEVID
- **Research grant DAAD (German Academic Exchange Service), 2017** final year project – FIBEVID
- **3rd best project SEECS Open House**, **2017**. Department-wide project competition at NUST – FIBEVID
- **2nd best project at COMPPEC, EME college**, **2017**. Nation-wide software project competition.
- **3rd best project at NASCON'17, FAST University, 2017**. Intra-city software project competition.

SELECT BLOGS

- **Nothing but NumPy: Understanding & Creating Neural Networks with Computational Graphs from Scratch**
  - 50K+ views; KDnuggests Gold Award
- **Nothing but NumPy: Understanding & Creating Binary Classification Neural Networks with Computational Graphs from Scratch**
  - 48K+ views; Published in Towards Data Science
- **Where did the Binary Cross-Entropy Loss Function come from?**
- **Why Using Mean Squared Error (MSE) Cost Function for Binary Classification is a Bad Idea?**
- **How do TensorFlow and Keras implement Binary Classification and the Binary Cross-Entropy function?**