


Muhammad Zohaib Nasir

+92 (301) 881 2374 | m.zohaibnasir6@gmail.com | github.com/mzohaibnasir | linkedin.com/in/m-zohaibnasir |  inthediary

SUMMARY

Senior ML/AI Engineering Leader with 6+ years driving enterprise AI transformation across Fortune 500 clients and high-growth organizations. Led cross-functional teams of 8+ engineers delivering revenue-generating AI products deployed at tier-1 enterprises (Jazz, Meezan Bank, Zong). Architected and scaled production systems from concept to deployment: cashierless retail infrastructure processing real-time transactions, autonomous multi-agent frameworks managing executive-level business operations, and biometric verification platforms serving millions of users. Spearheaded presales strategy—crafting technical proposals, building high-impact PoCs, and closing multi-million dollar deals. Domain expertise spans computer vision, generative AI, and multi-agent orchestration with proven ability to translate complex AI capabilities into measurable business outcomes. Track record of aligning AI initiatives with organizational strategy while maintaining technical excellence in production ML systems.

EDUCATION

Bachelor of Science in Computer Science

FAST National University of Computer and Emerging Sciences

Pakistan

EXPERIENCE

Software Alliance – Senior Machine Learning Engineer & Team Lead | Onsite

Jun 2024 – Present

- Lead the design and deployment of cashierless retail systems integrating real-time object detection, sensor fusion, and automation — enhancing operational efficiency and customer experience.
- Played a key role in presales: crafting proposals, developing PoCs, closing leads, and managing direct client relationships.
- Manage and mentor a team of 5 ML engineers, delivering 5+ successful AI projects across B2B and B2C sectors with measurable business impact.
- Architect scalable end-to-end ML systems - from data pipeline automation and model optimization to CI/CD-based deployment and real-time monitoring.
- Drive presales and client engagement, leading technical proposals, proof-of-concepts, and negotiations that contributed to closing high-value deals.
- Spearhead the integration and optimization of Large Language Models (LLMs) for enterprise-grade AI applications.
- Collaborate with leadership and cross-functional teams to align AI strategies with organizational goals and deliver innovative solutions in CV, NLP, and Generative AI.

AllZone Technologies – Senior Machine Learning Engineer & Team Lead | Onsite

Feb 2025 – Sep 2025

- Led a team of 5+ ML engineers, delivering 5+ successful AI projects across B2B and B2C domains with measurable business impact and strong client satisfaction.
- Architected and deployed end-to-end ML systems covering data pipelines, automated feature engineering, model evaluation, and production monitoring.
- Spearheaded enterprise-scale AI initiatives in Computer Vision, Natural Language Processing, and Generative AI, aligning technical innovation with strategic goals.
- Implemented CI/CD pipelines to streamline ML workflows, accelerate experimentation, and enable scalable deployment.
- Actively contributed to presales by drafting proposals, building PoCs, closing leads, and managing direct client relationships.
- Collaborated with cross-functional teams and stakeholders to translate business objectives into actionable AI strategies, ensuring seamless execution from concept to deployment.

AKSA-SDS – Senior Engineer (Artificial Intelligence) - Team Lead | Onsite

Jan 2022 – May 2024

- Engineered the Fleck-BVS biometric verification system with four-finger identification, deployed successfully at Jazz and Meezan Bank.
- Designed and implemented GAN-based data synthesis projects (DCGAN, WGAN, WGAN-GP) to generate high-quality synthetic training datasets.
- Enhanced KYC compliance by integrating advanced AI techniques, bolstering security and meeting strict regulatory standards.
- Earned a promotion to Senior Engineer within one year, demonstrating exceptional performance and leadership.
- Managed and mentored a team of five engineers, overseeing project execution and delivering high-impact solutions.

COMSATS – Research Assistant (AI) | Onsite

Jun 2023 – Dec 2023

- Contributed to the Fleck-BVS project under Dr. Mohsin Riaz, advancing biometric analysis with deep learning and digital image processing techniques.
- Enhanced image feature extraction and quality to support high-precision biometric verification.
- Optimized image processing workflows, achieving significant performance improvements in computer vision applications.

Althea AI– AllZone Technologies | *AI-Powered Medical Call Representative*

Feb 2025 – Sept 2025

- Customized Ultravox to autonomously manage medical clinic calls, including appointment scheduling, patient verification, and general inquiries.
- Curated domain-specific medical datasets and crafted task-oriented dialogue flows to train the assistant for high-precision interactions.
- Fine-tuned Ultravox's LLM backbone using QLoRA with Unsloth, Hugging Face PEFT for parameter-efficient adaptation; utilized Distributed Data Parallel (DDP) for scalable multi-GPU training.
- Integrated real-time API calls for appointment booking, patient verification, and clinical data validation to automate workflows.
- Deployed the system on Google Cloud Platform (GCP) with optimizations for low-latency, high-throughput inference under high call volumes.
- Impact: Reduced administrative overhead, improved patient engagement, and enabled cost-effective scalability for healthcare providers.

Genesys & LabelRx – AllZone Technologies | *Multi-Agent Automation Infrastructure*

Apr 2025 – Present

- Designed and implemented an agentic workflow framework with a central COO agent acting as the primary dispatcher and decision-maker for incoming tasks across multiple domains (GeneSys, LabelRx, Legal).
- Developed intelligent classification logic to categorize queries and annotate them with metadata, enabling accurate routing to downstream domain-specific agents (CMO, CFO, Legal, etc.).
- Constructed autonomous agents with role-specific capabilities: CMO for marketing automation, CFO for financial analytics, Legal for compliance, and COO for operations oversight and coordination.
- Enabled execution workflows within each agent to handle complex tasks such as financial forecasting, lead campaign generation, compliance validation, and regulatory risk analysis.
- Built a centralized aggregation and threshold evaluation layer that flags anomalies (e.g., Stripe revenue drops, marketing inefficiencies) and generates intelligent alerts for CEO-level intervention.
- Designed a real-time alerting mechanism that escalates strategic risks to the CEO (man-in-loop) with contextual summaries, consequence modeling, and actionable insights.
- Impact: Delivered a scalable multi-agent orchestration system that enhances decision accuracy, automates domain-specific operations, and enables proactive executive oversight.

Unlimit – Software Alliance | *AI + IoT Powered Cashierless Retail Experience*

Mar 2024 – Present

- Built an AI + IoT-powered cashierless checkout system allowing customers to grab items and leave—no scanning or waiting required.
- Implemented real-time object detection/segmentation and Multi-View Tracking (MVT) to accurately track customer-product interactions.
- Built a custom Re-Identification (ReID) pipeline integrating appearance-based feature extraction and cross-camera identity matching for robust multi-view tracking.
- Integrated weight sensors and load cells to verify pickups and drop-offs, significantly reducing false positives and maintaining accurate inventory.
- Resolved inference delays and frame rate issues over RTSP streams, ensuring smooth and consistent video performance.
- Automated end-to-end annotation and training pipelines for detection and segmentation using one-click scripts that ingest 3D object models, streamlining dataset creation and model iteration.
- Deployed the solution on AWS to ensure scalable real-time processing, analytics, and monitoring.
- Addressed real-world challenges such as occlusions, sensor desynchronization, and system latency to ensure reliability and performance.
- Tech Stack: Python, GStreamer, MLFlow, PyTorch, AWS, CUDA, Git, Docker, IoT.

Allah-u-Alim – Software Alliance | *RAG-Based Islamic Query Platform*

Aug 2024 - Feb 2025

- Developed an AI-powered platform delivering scholar-validated answers to Islamic queries, rooted in the Quran, Hadith, and traditional teachings for authentic, user-friendly guidance.
- Implemented contextual retrieval with BM25/TF-IDF algorithms, and integrated Anthropic's Claude 3-Haiku to generate contextually accurate responses.
- Optimized system performance through prompt caching to reduce computational costs and improve response efficiency.
- Utilized LangChain and LlamaIndex to build a robust RAG framework, leveraging open-source LLMs for advanced natural language understanding and generation.
- Improved information retrieval by fine-tuning embedding models, ensuring precise and relevant responses to nuanced theological queries.
- Engineered a comprehensive data pipeline for efficient retrieval and processing, enhancing the system's accuracy and response relevance.
- Deployed the platform on AWS for scalability, ensuring high availability and seamless user access.
- Collaborated with Islamic scholars to validate the doctrinal accuracy and authenticity of the platform's responses.
- Tech Stack: Python, AWS, LangChain, LlamaIndex, open-source LLMs, embedding models, Git.

- Developed an end-to-end automation system that processes emails, extracts order details, and autonomously manages sales and purchase order creation in Jira.
- Built a real-time email processing system using the Gmail API for automated email retrieval and content extraction.
- Automated Jira ticket creation for sales and purchase orders based on extracted email content.
- Integrated Google Drive API to manage and attach relevant documents to Jira tickets.
- Implemented order completion prediction using Google Calendar availability.
- Developed an automated Jira commenting system for real-time order tracking and updates.
- Deployed a FastAPI-based backend on AWS, ensuring scalability and high availability.
- Enabled seamless integration across Gmail, Jira, Google Drive, and Google Calendar for a fully automated system.
- Impact: Eliminated manual order processing, reduced errors and delays, and significantly improved workflow efficiency and productivity.

VLM-Based Image Analysis – Software Alliance | *Q/A, Segmentation, Detection*

Dec 2024 – Feb 2025

- Developed a visual question-answering system and an automated annotation tool using advanced vision-language models (VLMs) to enhance computer vision workflows.
- Optimized model performance by implementing key-value (KV) caching for efficient memory usage and faster inference during prompt-based tasks.
- Integrated SigLip as the image encoder to enhance feature extraction performance and support high-accuracy vision tasks.
- Deployed models using AWS EC2 and CUDA for accelerated training and inference.
- Tech Stack: PyTorch, AWS EC2, CUDA, Git.

Qwen2-VL – Software Alliance | *Quantization and Fine-Tuning*

Oct 2024 - Jan 2025

- Quantized and fine-tuned the Qwen2-VL model on a custom dataset, optimizing it for multimodal instruction-following tasks.
- Implemented Low-Rank Adaptation (LoRA) techniques to enhance model performance with minimal computational overhead.
- Designed and preprocessed the custom dataset, ensuring alignment with task-specific objectives for effective model training.
- Streamlined model deployment and inference pipelines, enabling efficient utilization on edge devices.
- Tech stack: Python, PyTorch, Hugging Face Transformers, LoRA, CUDA, AWS EC2, Git.

Fleck - Contactless BVS – AKSA-SDS | *Four-Finger Biometric Verification System*

Aug 2022 – Apr 2024

- Built a contactless four-finger biometric verification system that allows customers to verify their identity using mobile cameras in real-time, enhancing security for financial transactions.
- Applied contrast enhancement, noise reduction, and edge detection techniques to enhance the captured image's quality.
- Developed custom segmentation models and deep learning-based minutiae extraction techniques for precise fingerprint processing.
- Implemented minutiae-based matching and machine learning models for high-accuracy biometric verification.
- Deployed via Flask/FastAPI-based RESTful APIs, integrating with Jazz and Meezan Bank for real-time authentication.
- Established MLOps pipelines with continuous monitoring, version control, and automated model training for enhanced reliability.
- Tackled image quality variations due to environmental factors and implemented anti-spoofing measures.

LexiUS – AKSA-SDS | *AI-Driven Legal Assistant for American Law*

Feb 2024 – Mar 2024

- Created an AI-powered legal assistant providing expert guidance on American laws, legal precedents, and statutory interpretation for professionals and individuals.
- Built a Retrieval-Augmented Generation (RAG) pipeline for context-aware legal question answering using open-source LLaMA 3 for cost-efficient, customizable inference.
- Applied BM25 and TF-IDF algorithms for high-precision retrieval of relevant legal documents.
- Integrated LangChain for structured indexing and retrieval of case law, statutes, and legal commentary.
- Fine-tuned embedding models to enhance semantic understanding and document relevance.
- Developed data pipelines for preprocessing and structuring large volumes of complex legal documents.
- Collaborated with legal professionals to validate outputs, ensuring factual accuracy and compliance.
- Deployed models locally using AWS and Ollama for scalable, low-latency inference with reduced infrastructure costs.
- Implemented prompt response and API caching to optimize performance and minimize redundant computations.

- Developed a GAN-based biometric data synthesis pipeline to address data scarcity in biometric model training, enhancing model diversity while preserving privacy.
- Developed and implemented a DCGAN architecture to generate realistic synthetic biometric data, enhancing data diversity for model training.
- Utilized advanced GAN techniques, including WGAN and WGAN-GP, to significantly improve the quality and realism of the generated data.
- Tech stack: Python, FastAPI, MLFlow, PyTorch, CUDA.

Research Agent – RAG with Custom Agents

Jan 2024 - Jan 2024

- Created an end-to-end Retrieval-Augmented Generation (RAG) system using custom tools and retrieval chains, enhancing information retrieval capabilities.
- Leveraged only open-source models, promoting accessibility and collaboration.
- Utilized the Groq inference engine for high-performance, low-latency processing, ensuring efficient model execution.
- Tech stack: LLM (Gemma), Python, Streamlit, CUDA, Git, Langchain, Langsmith, Langserve, FAISS, GroqCloud.

KnowFace – AKSA-SDS | Onboarding Solution (KYC)

Apr 2022 – Aug 2023

- Developed a real-time AI-powered identity verification platform adopted by Zong and integrated with PayMax to streamline digital onboarding and ensure full KYC compliance.
- Implemented advanced facial recognition with liveness detection and spoof prevention for secure authentication.
- Built custom algorithms for precise face extraction and photo-ID matching.
- Integrated OCR (Tesseract) for automated data extraction from identity documents, including MRZ and barcode scanning.
- Developed RESTful APIs using Flask and FastAPI, supporting Android/iOS SDKs for easy integration.
- Applied MLOps principles, including model versioning and registry using MLFlow, for scalable deployment and maintenance.
- Tech stack: Python, Flask, FastAPI, MLFlow, PyTorch, AWS EC2, CUDA, Tesseract, Git.

Lexpal – Software Alliance | Predictive Typing & Multilingual Translation

Feb 2025 – Mar 2025

- Developed an AI-powered text assistant capable of next-word prediction and real-time translation across 50+ languages.
- Implemented statistical NLP techniques and language models for intelligent, context-aware text suggestions.
- Engineered an on-device translation module for fast, secure, and offline-capable multilingual support.
- Built a FastAPI backend to enable low-latency inference and seamless real-time interaction.
- Deployed the solution on AWS to ensure scalability, high availability, and performance under diverse workloads.
- Impact: Enabled seamless communication with fast, private, and reliable multilingual support—enhancing user experience and reducing reliance on external APIs.

Skills

Programming	Python, C/C++, Bash, CUDA
Web Frameworks	Flask, FastAPI, RESTful APIs, SDK Development (Android/iOS)
ML/DL Frameworks	PyTorch, Scikit-learn, Hugging Face (Transformers, PEFT), Unsloth, OpenVINO, DLStreamer
LLMs & Multimodal Models	LLaMA, Mistral, Phi-2, Falcon, Qwen, Qwen2-VL, Ultravox, Claude, GPT, Gemma, SigLip
Generative AI	RAG, LoRA/QLoRA, Unsloth, Chatbots, Vision-Language Models, VAEs, Stable Diffusion
NLP & Retrieval	GANs (DCGAN, WGAN, WGAN-GP), Vision Transformers, SAM, Data Synthesis
	BERT, GPT, NER, Summarization, Translation, Text Generation, BM25, TF-IDF
	Semantic Search, Contextual Retrieval, LangChain, LlamaIndex, LangGraph, LangSmith
	LangServe, FAISS, Pinecone, ChromaDB
Prompt Engineering	Few-shot, Chain-of-Thought, Function Calling, Prompt Caching, KV Caching
Computer Vision	YOLO, Faster R-CNN, Mask R-CNN, DeepLabv3, OpenCV, Detectron2, Albumentations
	GStreamer, VQA, Liveness Detection, Anti-Spoofing, OCR (Tesseract), MRZ/Barcode Scanning
3D & Tracking	PyTorch3D, Open3D, SORT, Deep SORT, Multi-View Tracking (MVT), Re-Identification (ReID)
	SIFT, ORB, Pose Estimation, Siamese Networks, Minutiae Extraction
Multi-Agent & Automation	Agentic Workflows, LangGraph, Agent Orchestration, ReAct Agents, Multi-Agent Coordination
	Workflow Automation, Task Routing, Decision Logic, Alerting Systems
IoT & Real-Time Systems	Sensor Fusion, Load Cells, Weight Sensors, RTSP Streaming, Real-Time Processing
MLOps & Infrastructure	Git, Docker, MLflow, CI/CD, Model Versioning, Monitoring, Model Quantization
	DDP (Distributed Data Parallel), Edge Deployment, On-Device Inference, LangServe
	Ollama, llama.cpp, Groq
Cloud & Deployment	AWS (EC2, S3, Lambda, SageMaker, Bedrock), GCP, Azure
Databases & Storage	PostgreSQL, MySQL, MongoDB, FAISS, Pinecone, ChromaDB
Domain Expertise	Healthcare AI, Legal AI, Financial Systems, Biometric Verification, KYC/Compliance