

# Ahmad Shahzad Khawaja

**Phone:** (+1) 2363132538 (Mobile) | **Email:** [a.shahzad2000july@gmail.com](mailto:a.shahzad2000july@gmail.com) | **LinkedIn:** <https://www.linkedin.com/in/ahmadshahzad2000/> | **Github:** <https://github.com/Ahmadshahzad2> |

**Address:** Greater Toronto Area, Canada (Home)

## ABOUT MYSELF

As a results-driven **Machine Learning Engineer and Data Scientist**, I specialize in building **production-grade AI systems** across diverse industries including sports, healthcare, finance, and marketing. With a strong foundation in **deep learning, NLP, and computer vision**, I've architected scalable pipelines that integrate **large language models, real-time data streams, and edge deployments**. My expertise spans the full ML lifecycle—from **model training and evaluation to cloud-native deployment** using modern infrastructure and MLOps tools. I take pride in translating complex AI technologies into solutions that drive measurable business impact, such as automating workflows, enhancing analytics, and accelerating decision-making. Whether collaborating in fast-paced teams or leading technical initiatives, I bring a high level of technical rigor, innovation, and execution.

## WORK EXPERIENCE

### NEXENTRIA – LITHUANIA

**DATA SCIENTIST** – 01/09/2024 – 04/2025

1. Developed an automated medical call transcription system for a pharmaceutical client, leveraging **open-source** and **third-party speech-to-text models** like **Deepgram** to transcribe complex medical jargon accurately.
2. Benchmarked transcription models using custom metrics such as **keyword error rate** and developed an **LLM-based approach** using **LangChain** to enhance transcription accuracy by correcting domain-specific terminology.
3. Implemented a secure process to redact personal information from call transcripts in compliance with **HIPAA** regulations, ensuring data privacy.
4. Enhanced customer service efficiency by automating form population, prioritization, and operator assignment, reducing manual workloads and response times.
5. Orchestrated deployment of the transcription service on **AWS**, utilizing services such as **Lambda, SageMaker**, and **Bedrock**, creating a real-time pipeline that streamlined the query resolution process.

### ATHENA AI

**MACHINE LEARNING ENGINEER** – 05/2023 – 09/2024

1. Designed and deployed an **end-to-end AI call analytics platform** for a digital marketing agency using **Deepgram, OpenAI GPT-4, AWS Lambda, SQS, Google BigQuery**, and **Looker Studio**. Automated transcription, classification, and dashboard generation, reducing manual call-tagging time by **70%** and achieving **90%+ classification accuracy**.
2. Built a **LangGraph-powered AI assistant** for a sports analytics team that integrated **PostgreSQL, Kafka, MinIO**, and **Milvus** to answer player performance queries in natural language. Utilized **GPT-4o, Amazon Bedrock**, and **NeMo Guardrails** for secure, real-time analysis and insight generation, enabling sub-5-second responses with 100% accuracy on technique-based queries.
3. Spearheaded diverse AI and machine learning projects, delivering cutting-edge solutions across industries including sports, security, finance, and education.
4. Developed **multi-media** AI models to classify advertisement content, integrating audio and video data for accurate categorization with a 93% F1 score.
5. Engineered a basketball action classifier leveraging **ConvLSTM** and **YOLOv8** models to provide professional-grade analytics for coaches, achieving 97% accuracy in shot detection.
6. Designed and deployed a real-time AI pipeline for predatory animal classification, integrating YOLOv8 with **NVIDIA DeepStream** and **AWS Lambda** for scalable deployment.
7. Created a **crypto trading chatbot** with natural language capabilities, integrating sentiment analysis, price predictions, and real-time trading execution.
8. Implemented 3D point cloud segmentation models for custom datasets, deployed via Flask APIs on EC2 instances.
9. Contributed to a SAAS application for building damage assessment, using YOLOv8-m for crack segmentation and categorization of structural issues from high-resolution drone footage.
10. Built an **AI-aided exam application** using GPT-3.5 turbo API, enabling efficient Q&A and evaluation from PDF documents, tailored for students and professionals.

### OMNO AI

**MACHINE LEARNING ENGINEER** – 05/2022 – 04/2023

1. Deploy and train machine and deep learning models to provide end-to-end vision and AI solutions.
2. Create pipelines to integrate deep learning models with non-AI solutions to satisfy all customer KPI's.

3. Conduct R&D and analyse existing literature study to identify state-of-the-art techniques to optimize solutions with the highest performance.
4. Successfully delivered 2 short-term and 2 long-term projects in 6 months

## EDUCATION AND TRAINING

---

09/2023 – 04/2025 Kamloops

**MASTERS IN DATA SCIENCE** Thompson River University

---

09/2018 – 05/2022

**BS ELECTRICAL ENGINEERING** Lahore University of Management Sciences

---

*Relevant Courses:* Computer Vision, Deep Learning, Machine Learning, Data Science, Digital Signal Processing, Communication Systems, Signal and Systems, Feedback Control Systems, Renewable Energy Systems, Electromechanical Systems, Electromagnetic Field and waves, Applied Probability and Statistics

**Final grade** 3.3/4

## SKILLS

---

### Technical Skills

---

#### Programming & Scripting

Python (advanced), Bash, R (intermediate), SQL

#### Machine Learning & Deep Learning

Supervised/Unsupervised Learning, Time Series Forecasting, Predictive Modeling, Model Evaluation (F1, AUC, KER)

Frameworks: **PyTorch**, **TensorFlow**, **scikit-learn**, **XGBoost**, **LightGBM**

#### NLP & LLM Integration

Prompt Engineering, LangGraph, LangChain, Retrieval-Augmented Generation (RAG), LLaMA Index, GPT APIs (3.5, 4, 4o)

Speech Processing: Deepgram (STT), ElevenLabs (TTS), Voice Cloning, NeMo Guardrails

#### Computer Vision

Object Detection, Segmentation, Action Recognition, Pose Estimation

Models: **YOLOv8**, **ConvLSTM**, OpenCV, MediaPipe

3D Vision: Point Cloud Segmentation, Depth Estimation

#### Data Engineering & MLOps

Cloud: **AWS** (Lambda, EC2, S3, SageMaker, Bedrock, EventBridge), GCP (BigQuery, Looker Studio)

Streaming & Messaging: Kafka, SQS

Containerization: Docker, Kubernetes

CI/CD: GitHub Actions, Jenkins

Monitoring & Logging: CloudWatch, custom pipelines

#### APIs & Back-End Systems

RESTful APIs, **Flask**, **FastAPI**, boto3, Webhooks

Database Systems: **PostgreSQL**, **MySQL**, **MongoDB**, **DynamoDB**

#### Tools & Platforms

Git, GitHub, Linux CLI, Jupyter, Looker Studio, Milvus, Redis, MinIO, Tavily

## PROJECTS

---

### AI Assistant for Real-Time Basketball Analytics

---

*LangGraph-powered multi-agent system for unified querying across stat feeds, game video, and performance metrics*

- Developed a **LangGraph-based chatbot** capable of interpreting natural language questions about player performance and routing them through specialized analytic paths.
- Integrated heterogeneous data sources:
  - **PostgreSQL** for user sessions and static data.
  - **MinIO** for shot-level event data and game clips.
  - **Kafka** for real-time statistical event streams.
  - **Milvus** for **semantic search** over technical basketball concepts and documents.
- Designed a **three-path intelligent router**:
  - **Metrics Path**: Code interpreter retrieves precise in-game metrics (e.g., average PalmAngle) from **MinIO**.
  - **Statistics Path**: Stream analytics module fetches game highs from **Kafka**.
  - **RAG Path**: **Milvus** + **GPT-4o** handles conceptual or strategy-based queries using semantic retrieval.
- Used **Amazon Bedrock agents** for orchestrating tool-augmented reasoning and **NeMo Guardrails** to ensure safe, relevant, and context-bound responses.
- Prompt engineering optimized for:
  - Context preservation
  - Live pseudocode execution

- Interpretability with debugging trace views
- Deployed for cloud environments (**AWS**, optionally **GCP**) with modular design enabling expansion to wearable data or team dashboards.
- Achieved:
  - **100% accuracy** on individual technique-based queries
  - **<5 seconds** average response time
  - **70% faster** metric extraction from unstructured sources

## AI Call Classification and Reporting Platform for PlumbersSEO

---

*serverless NLP-based call analytics system to automate transcription, classification, and reporting across thousands of service calls*

- Built a **scalable serverless pipeline** for fully automating the classification and reporting of inbound/outbound calls for a large digital marketing agency managing plumbing, HVAC, and electrical contractors.
- **Data Ingestion & Integration:**
  - Ingested audio via **CallRail API** and transcribed using **Deepgram**.
  - All incoming data, including call metadata, streamed through **Amazon SQS** to decouple ingestion from processing.
- **AI Classification Pipeline:**
  - Used **OpenAI GPT-4o** with a custom **prompt-engineering loop** to label each call by **service line**, **caller intent**, and **lead quality**.
  - Achieved **90%+ classification accuracy** through iterative prompt tuning and domain adaptation.
- **Data Storage:**
  - Call transcripts and tags stored in **Google BigQuery** for analytics; key summary fields cached in **DynamoDB** for low-latency retrieval.
- **Automated Reporting:**
  - Created dynamic client dashboards using **Looker Studio**, with daily refreshes orchestrated via **AWS EventBridge** and **Lambda** functions.
  - Admin overview reports aggregate metrics across all clients.
- **Deployment & Infrastructure:**
  - Built on a **fully serverless AWS architecture** (Lambda, SQS, EventBridge).
  - Framework designed to elastically scale with zero manual effort for onboarding new clients.
- **Impact Achieved:**
  - Reduced manual tagging time by **90%**.
  - Improved visibility and campaign ROI with real-time access to classified calls and insights.
  - Supported seamless reporting for hundreds of clients with automated nightly updates.

## AI Advertisement Content Classification

---

1. Developed a multi-media AI model for a sports advertisement company to classify videos into **content-advertisement** or **coverage-advertisement** categories. The model combined audio and image data from videos to create a unified encoding, which was classified by a machine learning model.
2. Achieved **93% F1 score** on testing data with high precision.
3. Deployed the system on **AWS Lambda** using EFS memory for scalable processing. Classified video metadata was stored in **S3**, and the URL was returned via an integrated **EC2 backend** using boto3

## Basketball Action Classifier

---

1. Created a custom dataset annotated temporally and spatially for **shot recognition** and **ball/player detection** models.
2. Trained **ConvLSTM models** to detect shots with **97% accuracy** and a **95% F1 score**, and classify phases like low-pocket and release with **90% accuracy** and F1.
3. Developed custom **YOLOv8 models** for ball and player detection, utilizing **Mediapipe** for pose landmarks and **ByteTrack** for tracking.
4. Implemented filtering logic to handle multiple players by tracking **ball-hand overlaps** across frames, ensuring focus on the main player.
5. Designed for professional basketball coaches to analyze player posture through **video clips** and **landmark-based analytics**.
6. Generated **metadata reports** with classified frame sequences, player information, and pose insights.

## Predatory Animal Classification

---

1. Built an AI pipeline to detect and classify predatory animals such as **foxes, mountain lions, dogs, cats, and humans** in backyards using CCTV images and motion sensors.
2. Used **Yolov8** for detection and classification. Deployed on **AWS Lambda** and integrated on **Jetson Nano** using **TensorRT** in NVIDIA DeepStream.
3. The system differentiated between wild and pet animals, providing phone notifications or triggering a house alarm based on the detected threat level.

## Crypto Trading Chatbot

---

1. Developed a chatbot capable of executing buy and sell crypto orders through natural language communication.
2. Incorporated detailed analysis of coin performance to guide users in their trading decisions.
3. Integrated sentiment analysis from ongoing news related to a coin to provide evidence-based investment advice.
4. Incorporated future price predictions to help users estimate potential profits.
5. Leveraged LLM's prompt engineering, NLP, and real-time crypto exchange integration for comprehensive trading support.
6. Backend development in Python Flask and frontend in Next.js.

## Building Damage Assessment

---

- Assisted in developing a SAAS application to generate health reports for old factories, buildings, and underground tunnels.
- Utilized the latest YOLOv8-m detector for crack segmentation on 4K drone footage.
- Classified building damage into distinct categories: cracks, blisters, and crazing.
- Implemented a sliding window training and inference process to maintain image quality and boost accuracy.
- Deployed the model on AWS EC2 using Flask with a front-end developed in React.

## AI-Aided exam

---

1. Developed an app using GPT-3.5 turbo API to create a chatbot capable of answering Q&A from a PDF with answers and runtime evaluations
2. Leveraged advanced prompt engineering and text comprehension techniques to build the app
3. Designed the chatbot to be ideal for students and professionals seeking to efficiently extract information from PDFs
4. Demonstrated proficiency in GPT-3.5 turbo API and its integration with app development

## Soccer Match Event Detection and Recognition

---

1. Developed a system to detect and recognize significant occurrences in soccer matches from broadcast feeds using machine learning techniques.
2. Trained and tested 3D models including ResNet 3D, SlowFast, Uniformer-V2, and Conv-LSTM to identify goals, shots on target, and shots off target in soccer matches.
3. Integrated a heuristic algorithm based on ball, player, and goal post detection to assist the 3D model in decision-making.
4. Achieved an F1 score of 70% across three classes using context-optimized frames of single events.
5. Utilized Python, TensorFlow, and Keras to implement the system.
6. Successfully applied the system to real-world soccer match videos.