MOHAMMED NIHAL KHAN

AI ENGINEER

+1(437) 556-9983 ◇ Ontario, Canada
mohammednihal281001@gmail.com ◇ LinkedIn ◇ Github ◇ Tech Portfolio

PROFESSIONAL SUMMARY

Results driven AI Engineer with 3+ years of experience specializing in AI/ML systems and production-ready solutions. Proven expertise in developing and deploying AI-powered applications usingPython, PyTorch, and cloud platforms. Strong background in MLOps, vector databases, and integratingLLMs from OpenAI and Anthropic into scalable production systems. Demonstrated ability to collaboratecross-functionally in agile environments while maintaining high code quality and continuous learningmindset

TECHNICAL SKILLS

Programming Languages:	Python, SQL, JavaScript.
AI/ML Frameworks:	PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, Lora, Qlora, Quantization
AI Technologies:	Large Language Model(LLMs), OpenAI API, Anthropic Claude API, Vector Databases
MLOps & Tools:	MLflow, Weights & Biases, Docker, Kubernetes, Git, CI/CD Pipelines
Cloud Platforms:	AWS (SageMaker, Lambda, S3, EC2), Azure ML, Google Cloud Platform
Databases:	PostgreSQL, MongoDB, Pinecone, ChromaDB, FAISS
Data Engineering:	Pandas, NumPy, Apache Airflow, ETL Pipelines, Data Preprocessing

EXPERIENCE

AI Engineer March 2025 – Present
NexApproach *Toronto, ON*

- Developed and deployed 5+ production-ready AI features using PyTorch and TensorFlow, improving user engagement by 35% and reducing processing time by 40%
- Implemented comprehensive AI performance metrics and evaluation frameworks that enabled data-driven prioritization of model improvements, resulting in 25% increase in model accuracy
- Architected and maintained MLOps pipelines using Docker, Kubernetes, and AWS SageMaker for seamless model deployment and monitoring of 10+ AI models in production
- Integrated OpenAI GPT-4 and Anthropic Claude APIs into customer-facing applications, processing 20K+ API calls monthly with 99.9% uptime
- Designed and implemented vector database solutions using Pinecone and ChromaDB for semantic search functionality, improving search relevance by 45%
- Collaborated with cross-functional teams of 8+ engineers and data scientists in agile sprints, participating in code reviews and delivering features on schedule

Software Engineer Jan 2024 - Jan 2025
Cell2Fix *Brampton, ON*

- Built and optimized machine learning pipelines for data processing and model training, reducing training time by 50% through efficient data preprocessing
- Implemented RESTful APIs using FastAPI for serving ML models, handling 10K+ requests daily with average response time under 200ms
- Developed automated data validation and quality checks for ML datasets, reducing data-related errors by 60%

- Collaborated with senior engineers to translate prototype models into production-ready code following best practices for maintainability and scalability

Research Assistant April 2020 - May 2023
Sathyabama University *Chennai, TN*

- Built a Python-based computer vision system that hooked directly into lab microscopes, which allowed us to identify different larval stages in real-time and replaced a very tedious manual process.
- Eliminating Human Error by using OpenCV to create live detection overlays, I gave researchers instant visual confirmation on their screens, which increased data-collection speed by 50% and virtually eliminated manual counting errors.
- Worked side-by-side with the biology team to refine our machine learning models, ensuring the software was sensitive enough to catch tiny physiological changes in the larvae that were easy to miss.

EDUCATION

M.Eng in Computer Science, University of Windsor 2023 - 2024
B.Tech in Information Technology, Sathyabama University 2019 - 2023

PROJECTS

Digital Twin Portfolio Chat (Gemini 2.5 Pro & RAG):

- Created a Digital Twin chat window inside my Node.js portfolio that answers recruiter questions as if it were me.
- Implemented a RAG pipeline using the Gemini 2.5 Pro API and Hugging face to query my personal technical documents, providing a seamless and personalized experience for over 100+ visitors.(Try it here)

FUT Stats Tracker (OCR and Advanced Data Structures):

- Developed a full-stack dashboard for FIFA players that eliminates manual entry by using an OCR scanner to read player cards dragged and dropped from sites like Futbin.
- Built an analytics engine that tracks match by match performance, utilizing advanced data structures to calculate complex metrics like top-scorers and best-in-position ratings. (Try it here)

Medical Imaging: Brain Tumor Analysis (Approved Patent):

- Designed a CNN-based pipeline (TensorFlow/Keras) to classify MRI scans as malignant or non-malignant with 92% accuracy.
- Used OpenCV and Albumentations to improve precision by 15% on low-resolution scans; project resulted in an approved technical patent.

Environmental AI: Fish Species Identification (PySpark and AWS):

- Engineered a large-scale classification pipeline on AWS EMR to identify 20+ fish species, utilizing adaptive histogram equalization to improve accuracy in low-light conditions.
- Automated metadata tagging using BERT and stored results in Snowflake for high-speed collaborative analytics.

ASL Hand Recognition (TensorFlow and Edge Computing):

- Trained a CNN on 5,000+ images to recognize American Sign Language signs, achieving 98% accuracy.
- Optimized the model for Raspberry Pi hardware, reaching a ¡5ms inference time for real-time edge use.