

MUZNA NAEEM

Machine Learning Engineer

Islamabad, Pakistan | +92 330 568 8778 | muznanaeem1999@gmail.com
linkedin.com/in/muznanaeem | muznanaeem26.github.io

Summary

Machine Learning Engineer who builds and ships production AI systems end-to-end, from LLM agents and generative AI to recommendation, search, and the MLOps that keeps them reliable. I turn research ideas into low-latency services that run at scale on Kubernetes and GCP, and I measure everything I build.

Education

Bachelor of Science in Electrical Engineering (Specialization in Computer Engineering)

FAST NUCES, Islamabad | 2022

Honors: Magna Cum Laude · Dean's List (All Semesters)

Experience

Scala Teams | Islamabad (Hybrid)

Machine Learning Engineer | Jun 2024 – Present

- Shipped 10+ production ML services end-to-end across generative AI, agentic workflows, recommendation, search, and data infrastructure for enterprise e-commerce, owning each from prototype to deployment on Kubernetes and GCP.
- Built an agentic catalog-automation platform (multi-agent orchestration with human-in-the-loop approval) that lets non-technical teams run catalog operations in plain language, cutting manual catalog workload by ~40%.
- Delivered a broad GenAI suite: a multimodal image-to-attribute generation service, RAG-based retrieval, and an LLM review summarizer that distills customer feedback at scale.
- Designed and deployed transformer-based recommendation (session-based, frequently-bought-together, alternate and visual product recommendations), neural semantic search, and learning-to-rank systems serving live production traffic at low latency.
- Built a dynamic pricing and price-optimization model to lift sales and a product popularity scoring service for ranking and merchandising, backed by event-driven data pipelines (Kafka, MongoDB, BigQuery) and an ML data-quality monitoring system.
- Partnered with product and engineering teams to translate business needs into reliable, production-ready AI features across the e-commerce stack.
- Established evaluation, documentation, and reproducibility standards across ML projects, improving maintainability and accelerating team onboarding.

UpStart Commerce | Islamabad (Hybrid)

Software Development Engineer in Test (SDET) | Jul 2023 – Jun 2024

- Built UI/API automation frameworks (Cypress, Java, REST APIs), improving regression reliability and CI/CD quality.
- Designed integration and regression testing strategies that reduced production defects.
- Partnered with engineering teams to debug distributed workflows and improve release stability.

FAST NUCES | Islamabad

Lab Instructor | Jul 2022 – Jun 2023

- Taught Machine Learning, Programming Fundamentals, Embedded Systems, and Digital Communication.
- Mentored students on ML pipelines, evaluation methodology, and reproducible experiments.

HigherSummit | Islamabad

AI Intern | Dec 2022 – Jan 2023

- Built NLP chatbot systems and supported data preprocessing and model-training pipelines.
- Assisted in neural-network experimentation and dataset preparation.

Skills

Programming: Python, SQL, C++, MATLAB, Bash, Linux, Git, GPU Programming

ML / AI: LLMs, Transformers, RAG, Fine-tuning, Prompt Engineering, Tool Calling, MCP, Multi-Agent Systems, NLP, Computer Vision, Deep Learning

Recommendation: Learning-to-Rank, Collaborative Filtering, Matrix Factorization, Sequential & Session-based Models

Search / Retrieval: Vector Search, Embeddings, Semantic Search, BM25, TF-IDF, FAISS, Qdrant, Elasticsearch

Frameworks: PyTorch, TensorFlow, HuggingFace, LangChain, LangGraph, Google ADK, XGBoost, Scikit-learn, Transformers4Rec, SentenceTransformers, NVTabular, Pandas, NumPy

MLOps: Docker, Kubernetes, CI/CD, GitOps, MLflow, GitHub Actions, Bitbucket Pipelines, NVIDIA Triton, FastAPI

Data Engineering: Kafka, MongoDB, BigQuery, Airflow, Event-driven architectures

Cloud: Google Cloud (Vertex AI, Cloud Run, Agent Engine), Azure

Monitoring / Eval: LangSmith, Arize AI, Phoenix, LLM-as-Judge, A/B Testing, Model Monitoring, Alerting

Projects

AI-Powered PIM Agentic System | Scala Teams

Tech: LangGraph, Google ADK, Gemini, MCP, Vertex AI, FastAPI, Kubernetes

- Built a conversational AI system for enterprise catalog management using multi-agent orchestration with human-in-the-loop approval.
- Cut manual catalog operations workload by ~40% by letting non-technical teams work in natural language.

Multimodal Product Generation System | Scala Teams

Tech: Vertex AI, Gemini, MongoDB, MLflow, FastAPI

- Built an image-to-structured-data GenAI pipeline that auto-generates product attributes for e-commerce catalogs.
- Cut manual data entry from ~15 minutes to ~30 seconds per product across 50K+ products.

Neural Search Fine-tuning & Evaluation System | Scala Teams

Tech: SentenceTransformers, PyTorch, FAISS, MNR Loss

- Fine-tuned embedding models to improve semantic search quality over a keyword baseline.
- Improved search relevance by +22% (nDCG@10), validated against the industry-standard ESCI benchmark.
- Designed a multi-metric evaluation framework (ranking quality, diversity, coverage) to make results defensible.

Session-Based Recommendation System | Scala Teams

Tech: Transformers4Rec, XLNet, NVTabular, FastAPI

- Built a sequential recommender that predicts the next item from a user's live browsing session.
- Improved next-item predictions by +31% (Recall@20) over a popularity baseline.
- Kept responses fast in production at under 80 ms (p95) inference latency.

Certifications

NLP with Classification & Vector Spaces (DeepLearning.AI) | Google Cloud Platform Fundamentals: Core Infrastructure | Foundations: Data, Data Everywhere (Google) | Programming for Everybody: Python (University of Michigan) | Exploratory Data Analysis with Seaborn | Unit Testing & TDD in Python (Udemy) | REST APIs with Postman (Udemy)