

MUNIM AHMAD

Full-Stack AI Engineer | AI Systems | Deployment Platforms

Lahore, Pakistan • munimahmad2@gmail.com • +92 316 4086827
munimahmad.me • linkedin.com/in/munimahmad • github.com/munimx

PROFILE

Full-Stack AI Engineer building developer platforms, cloud infrastructure, AI systems, and analytics workflows. Reduced inference latency 65%, shipped production RAG/NL-to-SQL pipelines, and delivered TypeScript/Python cloud-native tools.

PROFESSIONAL EXPERIENCE

Backend AI Engineering Intern | FlyRank AI - Lahore, Punjab, Pakistan (Remote) | June 2026 - Present

- Build backend AI services that orchestrate LLM-powered extraction, classification, and content-analysis workflows behind typed APIs and asynchronous job handlers.
- Implement validation, retry, and structured-output layers for AI pipelines that transform URL, ranking, and content signals into reliable backend artifacts.

AI/ML Engineering Intern | Developers Hub Corporation | April 2026 - May 2026

- Built a Telco Customer Churn prediction pipeline with scikit-learn ColumnTransformer preprocessing, Logistic Regression, Random Forest, GridSearchCV, F1/ROC-AUC evaluation, and joblib model export.
- Developed an NLP support-ticket tagger comparing BART zero-shot, FLAN-T5 few-shot, and DistilBERT fine-tuning; few-shot classification reached 76.7% accuracy, 0.729 F1, and 90% top-3 accuracy.

Full-Stack AI Engineer | Endshift - Lahore, Pakistan | May 2025 - January 2026

- Architected a production RAG and natural-language-to-SQL pipeline using local LLMs, ChromaDB, prompt evaluation, and semantic search, serving 500+ daily queries with 98% uptime.
- Built Power BI operations dashboards from PostgreSQL exports and 42K+ IPDR records, tracking query volume, SLA aging, model usage, support trends, DAX KPIs, and Power Query cleanup.
- Reduced Mistral-7B inference latency by 65% through batching, 4-bit quantization, caching, and model-serving optimization.

Full-Stack MERN Developer | Greyloops - Lahore, Pakistan | August 2024 - April 2025

- Built React, Node.js, Express, and MongoDB portal features including authenticated CRUD flows, reusable forms, dashboard widgets, search, filters, and MongoDB aggregation-backed reports.
- Developed Power BI dashboards from MongoDB/PostgreSQL exports and 38K+ support/order records, tracking ticket aging, active users, fulfillment status, feature usage, and weekly QA trends.
- Integrated REST APIs with pagination, validation, role checks, and error handling while resolving 40+ staging bugs across weekly releases.

TECHNICAL PROJECTS

Liftoff - Deploy-as-a-Service Platform | tryliftoff.tech | February 2025 - Present

- Built a multi-tenant deployment platform that provisions infrastructure directly into user-owned DigitalOcean accounts via Pulumi, handling per-user credential isolation, cross-account Docker image builds, DOCR pushes, and App Platform provisioning.
- Designed the Next.js and NestJS control plane for GitHub OAuth, project/environment/repository management, encrypted cloud credentials, Prisma deployment state, Redis/BullMQ queues, and real-time Socket.io deployment logs.
- Implemented a canvas graph compiler that converts services, resources, and connection nodes into deployable infrastructure specifications with automatic environment-variable injection.
- Built the platform deployment path on Azure Container Apps with ACR-backed images, GitHub OIDC, managed identity, and CI/CD for repeatable platform rollouts.

Tech: TypeScript, Next.js, NestJS, Prisma, PostgreSQL, Redis, Docker, Pulumi, DigitalOcean, Azure, GitHub Actions

Recallm - Python Semantic Cache Library for LLM Apps | recallm.dev | March 2026 - Present

- Built and published an open-source Python package (`pip install recallm`) that wraps OpenAI-compatible chat completions with semantic caching, achieving sub-10ms cache lookups.
- Implemented ONNX embeddings, cosine similarity search, context-hash safeguards, TTL, namespaces, Redis/in-memory backends, and observability, reducing LLM costs 40-70% on FAQ-style workloads.

Tech: Python, NumPy, FastEmbed, ONNX Runtime, SentenceTransformers, Redis, Prometheus, Grafana, Pytest, Ruff, MyPy

Snag - Webhook Inspector and AI Agent Bridge | snag-web-five.vercel.app | June 2026 - Present

- Built a Next.js/Fastify webhook console with CLI tunneling, WebSocket forwarding, replay, JSON diffing, BullMQ queues, Redis, and a Python MCP bridge for local AI-agent inspection.

EDUCATION

B.Sc. Computer Science | University of Central Punjab, Lahore | July 2026

Relevant Coursework: Machine Learning, Deep Learning, Parallel & Distributed Computing, Data Structures, Database Systems

TECHNICAL SKILLS

Languages: TypeScript, Python, C#, SQL

AI/ML: PyTorch, TensorFlow, Hugging Face, LangChain, LlamaIndex, OpenAI/Anthropic APIs, ChromaDB, FAISS, LoRA/QLoRA

Full-Stack: React, Next.js, NestJS, Node.js, Express, MongoDB, .NET, FastAPI, REST, GraphQL, WebSockets

Data & Cloud: Power BI, DAX, Power Query, PostgreSQL, Redis, Docker, Kubernetes, AWS, Azure, DigitalOcean