MUHAMMAD AHMED AFZAL

ahmed.afzal5114@gmail.com | +92-332-7724222 | Islamabad, Pakistan

LinkedIn: muhammad-ahmed-afzal

PROFESSIONAL SUMMARY

Production-focused AI Engineer with proven experience building and deploying scalable ML systems. Delivered high-impact solutions including AI APIs serving 50K+ daily requests with <200ms latency and enterprise automation platforms achieving 85% operational efficiency gains. Strong foundation in cloud infrastructure (GCP/AWS), MLOps pipelines, and production deployment. Specialized in PyTorch, FastAPI, transformer architectures, and building reliable AI systems from prototype to production. Seeking to leverage technical expertise and hands-on delivery experience in a growth-oriented AI engineering role.

TECHNICAL SKILLS

Core Stack: Python, PyTorch, TensorFlow, FastAPI, Docker, Git, SQL

Ml Ai: Transformers (BERT/GPT), Reinforcement Learning, Computer Vision, LangChain, Hugging Face

Cloud Devops: GCP (Cloud Run, App Engine), AWS (EC2, RDS, S3), CI/CD pipelines, monitoring

Data Apis: PostgreSQL, MySQL, MongoDB, Redis, REST APIs, GraphQL, WebSocket

Additional: JavaScript/TypeScript, Node.js, React, Flutter, Dart

PROFESSIONAL EXPERIENCE

AI Engineer | iTechGemini Pvt Ltd | Lahore, Pakistan

Sep 2025 – Present

- Built and deployed 5+ production AI APIs serving mobile applications with 99.7% uptime, integrating OpenRouter, Replicate, and Hugging Face models for real-time inference at scale.
- Engineered cloud-native microservices on GCP (Cloud Run, App Engine) optimizing for <200ms p99 latency while handling 50K+ daily API requests through auto-scaling and load balancing.
- Implemented MLOps pipelines including model versioning, A/B testing infrastructure, and Prometheus monitoring dashboards, reducing deployment cycles from 4 hours to 25 minutes.
- Optimized API performance through Redis caching, async request handling, and connection pooling achieving 3.2x throughput improvement and 40% cloud cost reduction.
- Developed multi-modal AI features spanning computer vision, NLP, and generative AI with fallback strategies and error handling.

AI Engineer | Automotive Artificial Intelligence (AAI) GmbH | Islamabad, Pakistan

Dec 2024 – Apr 2025

- Designed and deployed GoldenEYE—an ETL automation platform replacing manual data workflows, achieving 85% reduction in manual effort and enabling hourly data ingestion with zero downtime.
- Architected AWS RDS/EC2 MySQL infrastructure handling 100K+ records and 10K+ daily transactions with <50ms query latency through indexing optimization and connection pooling.
- Built real-time Python analytics dashboards with <300ms response time improving operational efficiency by 70% and enabling data-driven decision making across global teams.
- Implemented automated data validation pipelines using statistical anomaly detection improving data quality by 92% and reducing manual QA overhead.

- Established monitoring and alerting systems ensuring 99.9% platform availability.

AI Engineer | Freelance / Contract | Islamabad, Pakistan

Apr 2025 – Sep 2025

- Enhanced 3DmolFormer transformer model with multi-objective RL for protein-ligand binding prediction, achieving 18% improvement in binding affinity MAE (1.42 → 1.16 kcal/mol) and 0.89 ROC-AUC on PDBbind benchmark.
- Built full-stack luxury booking platform with React/Node.js/PostgreSQL featuring GPT-4 chatbot and ML recommendation engine, deployed on AWS with 99.9% availability serving 1K+ bookings.
- Developed Flutter sports social app with ML matchmaking algorithms and real-time Firebase integration, successfully launched on Play Store with 500+ active users.

Research Assistant | COMSATS University Islamabad | Islamabad, Pakistan

Sep 2022 – Jul 2023

- Conducted NLP research on transformer models (BERT, GPT-2) for sentiment analysis achieving 15% accuracy improvement through fine-tuning and hyperparameter optimization.
- Developed CNN architectures for medical image classification reaching 92% accuracy and 0.91 F1-score on benchmark datasets.
- Co-authored research on scalable deep learning architectures bridging academic innovation with production deployment.

EDUCATION

Master of Science in Artificial Intelligence - COMSATS University Islamabad (Feb 2024 – Jan 2026)

GPA: 3.50/4.00

Bachelor of Science in Bioinformatics - COMSATS University Islamabad (Sep 2019 – Jul 2023)

GPA: 2.82/4.00

KEY PROJECTS

3DMOLFORMER: Advanced transformer architecture for 3D molecular representation learning in drug discovery with dual-channel attention mechanism.

MolGAN: Generative Adversarial Network for de novo molecular design using reinforcement learning reward shaping.

Translation AI API: Production FastAPI service supporting 50+ languages with async handling, rate limiting, and <500ms inference time.

AI LinkedIn Agent: Automated networking bot using LLMs with RAG architecture for context-aware content generation.