Muhammad Ahmed Afzal | AI Engineer | Production ML Systems & Cloud Infrastructure

+92-332-7724222 • ahmed.afzal5114@gmail.com
muhammad-ahmed-afzal • AHmedaf123 • Islamabad, Pakistan

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

iTechGemini Pvt Ltd — Lahore, Pakistan
AI Engineer — *Sep 2025 – Present*

Key Achievement: Production AI API Infrastructure

- 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

Automotive Artificial Intelligence (AAI) GmbH — Islamabad, Pakistan
AI Engineer — *Dec 2024 – Apr 2025*

Key Achievement: GoldenEYE Automation System

- 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**

Freelance / Contract — Islamabad, Pakistan
AI Engineer — *Apr 2025 – Sep 2025*

Client Project Delivery:

- **Drug Discovery AI:** 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
- **SiteNest Platform:** 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
- **PlayAround Mobile App:** Developed Flutter sports social app with ML matchmaking algorithms and real-time Firebase integration, successfully launched on Play Store with 500+ active users

COMSATS University Islamabad — Islamabad, Pakistan
Research Assistant — *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

COMSATS University Islamabad — Islamabad, Pakistan
Master of Science in Artificial Intelligence, GPA: 3.50/4.00 — *Feb 2024 – Jan 2026*
Coursework: Advanced Deep Learning, Reinforcement Learning, Natural Language Processing, Computer Vision

COMSATS University Islamabad — Islamabad, Pakistan
Bachelor of Science in Bioinformatics, GPA: 2.82/4.00 — *Sep 2019 – Jul 2023*
Foundation: Machine Learning, Data Structures, Algorithms, Software Engineering, Statistical Analysis

Notable Projects

3DMOLFORMER: github.com/AHmedaf123/3DMOLFORMER – Advanced transformer architecture for 3D molecular representation learning in drug discovery with dual-channel attention mechanism

MolGAN: github.com/AHmedaf123/molgan – Generative Adversarial Network for de novo molecular design using reinforcement learning reward shaping

Translation AI API: github.com/AHmedaf123/Translation-AI-API – Production FastAPI service supporting 50+ languages with async handling, rate limiting, and <500ms inference time

AI LinkedIn Agent: github.com/AHmedaf123/ai-linkedin-agent – Automated networking bot using LLMs with RAG architecture for context-aware content generation

Healthcare Stroke Classification: github.com/AHmedaf123/Stroke_Classification – Multi-class stroke classification achieving 89% accuracy and 0.87 F1-score using ensemble deep learning

SAM Implementation: github.com/AHmedaf123/sam – Production integration of Segment Anything Model for real-time image segmentation with optimized inference pipeline

Technical Achievements

Delivered **8+ production ML systems** with enterprise-grade reliability (99.7%+ uptime) and performance optimization

Generated measurable business impact including **85% operational cost reduction** and **70% efficiency improvements**

Maintained **154+ open-source repositories** covering ML infrastructure, drug discovery AI, computer vision, and NLP

Specialized in low-latency inference optimization, high-throughput API design, and cloud-native ML deployment

Languages

English: Fluent – Professional technical communication and documentation

Urdu: Native