# Muhammad Abdul Mughni

+92-3134303345 mughniqureshi@gmail.com linkedin.com/in/m-abdul-mughni Portfolio

## Summary

AI Engineer with 2+ years of experience building production Generative AI systems from RAG-based chatbots with vector search to real-time computer vision apps. Cut inference latency by 44% (3.2s→1.8s), designed features that boosted engagement by 35%, and deployed scalable ML pipelines serving 10K+ daily requests. Skilled in LLM orchestration, fine-tuning, NLP, and end-to-end model deployment.

## Experience

**Associate AI Engineer** June 2024 – Present
*Terafort* *Onsite*

- **Generative AI Social App**
  * Architected GenAI features using LangChain, PyTorch & TensorFlow — built content moderation pipeline with safety guardrails reducing false positives by 50%, and engineered a RAG system with prompt chaining driving 35% increase in user engagement.
  * Optimized inference via quantization & batching, cutting latency from 3.2s→1.8s (44%) serving 10K+ daily requests with no quality loss.
- **AI Photo Editor — Picsly**
  * Engineered AI-powered photo enhancement pipelines for real-time image generation, and style transfer templates — shipped to **1M+ users** on Google Play with a 4.0 rating.
- **AI-powered Interior Design App**
  * Developed an AI-powered interior design system using computer vision and generative models, with image segmentation & object detection (YOLO/Mask R-CNN) for accurate furniture recognition, spatial adjustments, and dynamic style suggestions.
  * Built a real-time visualization module for high-quality AI-generated room renderings and optimized the inference pipeline, achieving 30% faster design previews.

## Projects

**Domain-Specific RAG Chatbot** | *LangChain, FAISS, FastAPI*

- Production-deployed domain expert chatbot with FAISS semantic search over 5K+ docs, improving relevance by 30%. Served via FastAPI + Docker with 90% satisfaction. Extending into autonomous AI agent workflows with tool-use and memory.

**AI-Powered Weather Briefing Bot** | *Python, AWS lambda, LLM*

- Built a serverless, multi-user Discord bot on AWS that aggregates real-time data, processes it through a large language model to generate personalized daily briefings, and delivers them to users on a scheduled basis. Implemented OAuth2 authentication, event-driven architecture, and infrastructure as code for automated cloud deployment.

## Education

**University of Engineering and Technology** 2020-2024
*Bachelor of Science in Computer Science* *Lahore, Pakistan*

## Technical Skills

**Expertise**: Machine and Deep Learning, Agentic AI, Fine-Tuning (LoRA/QLoRA), RAG, AI Automations, Prompt Engineering, Responsible AI, Technical Documentation
**AI/ML Frameworks**: Pytorch, Hugging Face, OpenAI, LangChain, LlamaIndex, Scikit-learn, OpenCV
**MLOps & Deployment**: Docker, Flask, FastAPI, AWS, GPU Cloud, MLflow, Linux
**Data & Tools**: NumPy, Pandas, Matplotlib, Jupyter, Git, n8n, Claude Code
**Soft Skills**: Innovative, Collaborative, Data-driven, Articulate, Adaptability, Analytical & Problem-solving skills

## Leadership & Achievements

- Data Science Lead at Google Developer Student Club (GDSC) - UET LHR
- Selected as Certified **GDSC Keynote Speaker.**