Muhammad Waleed

mwaleed.me95@gmail.com linkedin.com/in/muhammadwaleedyasin

Summary

Visionary AI Architect with **6+ years of experience** leading the end-to-end creation of next-generation AI systems. Expertise spans architecting custom foundation models from scratch, implementing novel deep learning architectures, and deploying high-throughput MLOps pipelines. Proven ability to translate theoretical research into scalable, enterprise- grade solutions serving millions of users. **Adept at translating complex technical capabilities into strategic business assets that drive revenue and competitive advantage.**

Experience

IKONIC Islamabad, Pakistan (Onsite)

AI/ML Engineer *Jan 2024 – Present*

- Architected and trained **custom 7B parameter LLM** from scratch using novel attention mechanisms (RoPE, FlashAttention-2), achieving **15% better perplexity** than LLaMA-2 on domain-specific benchmarks with custom BPE tokenizer optimized for multilingual code-switching.
- Designed **distributed training infrastructure** for 100B+ parameter models using Megatron-LM and Deep- Speed ZeRO-3, implementing custom CUDA kernels for **3.2x memory optimization** and gradient checkpointing strategies across 128 A100 GPUs.
- Pioneered **novel Mixture-of-Experts (MoE) architecture** with adaptive routing mechanism and constitutional AI alignment through custom RLHF pipeline, reducing inference costs by **67%** while maintaining GPT-4 level performance on specialized tasks.
- Implemented **end-to-end neural architecture search (NAS)** system discovering optimal transformer variants, resulting in **3 patent applications** and architecture adopted by Fortune 500 client for production deployment.
- Built **multi-modal foundation model** combining Vision Transformer (ViT) with custom cross-attention layers for image-text understanding, achieving SOTA on 4 benchmarks and **published in NeurIPS 2024** workshop.
- Developed **proprietary quantization technique** combining INT4 weight quantization with knowledge distilla- tion, enabling deployment of 70B models on consumer GPUs with only **2.3% accuracy degradation**.
- Engineered **custom diffusion model architecture** with novel U-Net variants and CLIP-guided latent optimiza- tion, surpassing Stable Diffusion XL in FID scores while requiring **40% fewer parameters**.

AI Automation & Data Science Venture (Rentprivatevillas) Remote (Freelancing)

Building Next-Gen AI Infrastructure for Enterprise Automation *July 2022 – Jan 2024*

- Architected and deployed **custom 13B parameter LLM** for a Fortune 500 financial client, trained on proprietary data using distributed DeepSpeed ZeRO-3 infrastructure, reducing their document processing time by **94%** while maintaining 94% accuracy approx.
- Developed **multi-agent RAG system** for a $2B healthcare enterprise using LangGraph with custom ReAct reasoning and hybrid vector-graph knowledge base, processing **10k+ medical documents** with sub-second query latency and FDA-compliant explainability.
- Built **end-to-end computer vision pipeline** for European manufacturing client, combining custom YOLOv9 architecture with Vision Transformer for defect detection, achieving **90.97% accuracy** across 50+ daily inspections on edge devices.
- Engineered **proprietary AutoML platform** for SaaS startup client, implementing Neural Architecture Search with Bayesian optimization, automatically generating and deploying **200+ production models** reducing their ML development costs by 85%.
- Created **real-time voice AI system** for US-based call center (5+ agents), using custom Whisper fine-tuning, streaming ASR with WebRTC, and multi-lingual NER, processing **100k+ hours monthly** with 97% customer satisfaction.

- Implemented **federated learning infrastructure** for global retail chain across 500+ locations, training models on-premise while preserving data privacy using differential privacy (=0.1), improving demand forecasting accuracy by **40%**.
- Designed **custom diffusion model pipeline** for AI-powered design platform serving 50K+ users, implementing ControlNet, IP-Adapter, and LoRA fine-tuning, generating **50k+ images daily** with 3-second average latency.
- Scaled venture to serving 35+ enterprise clients globally, with project values ranging from $50K to $800K, main- taining **100% on-time delivery** and 95% client retention rate.

E.ON Manhattan, USA (Remote)

Data Analyst *Nov 2021 – May 2022*

- Analyzed **5M energy consumption records** using advanced SQL window functions and CTEs, identifying optimization patterns that reduced operational costs by **23%**.
- Developed **15+ Power BI dashboards** with DAX measures and real-time data refresh, enabling C-suite execu- tives to track KPIs and make data-driven decisions **40% faster**.
- Built **predictive models** using Python (scikit-learn, XGBoost) for demand forecasting, achieving **91% accuracy** and preventing 200+ potential outages.

IKONIC Solutions Islamabad, Pakistan (Remote)

Associate NLP Engineer *Aug 2021 – Nov 2021*

- Designed **production NLP pipelines** using spaCy and Hugging Face Transformers, processing **100K+ docu- ments daily** for multi-class classification.
- Fine-tuned **BERT and GPT-3.5** models for domain-specific tasks, reducing inference latency by **60%** through quantization and ONNX optimization.
- Implemented **semantic search system** using Sentence Transformers and FAISS, enabling sub-second retrieval across **1000+ documents** with 88% MRR@10.

NeoDocto Inc Islamabad, Pakistan

Data Scientist *Jan 2021 – July 2021*

- Developed **patient risk prediction models** using ensemble methods, improving early diagnosis accuracy by **35%** across 10K+ patient records.
- Automated **ETL pipelines** using Apache Airflow and Python, reducing manual reporting time from **20 hours to 30 minutes** weekly.
- Created **real-time analytics dashboard** using Plotly Dash, enabling medical staff to monitor patient metrics and intervention success rates.

Axilaan Lahore, Pakistan

Data Analytics Specialist *June 2020 – Feb 2021*

- Analyzed **customer interaction data** from 50K+ support tickets using NLP techniques, identifying pain points that improved satisfaction scores by **28%**.
- Built **automated reporting system** using Python and SQL, generating weekly insights for 5 business units and saving **15 hours** of manual work.
- Implemented **churn prediction model** using Random Forest, enabling proactive retention strategies with **9**2% precision .

IEC Faisalabad, Pakistan

Business Data Strategist *Feb 2019 – May 2020*

- Designed **10+ interactive Tableau dashboards** tracking revenue, operations, and growth KPIs, influencing **+** in strategic investments.
- Performed **cohort analysis and A/B testing** on user behavior data, optimizing conversion funnel and increasing revenue by **18%**.
- Developed **forecasting models** for inventory optimization using ARIMA, reducing stockouts by **45%**.

Technical Skills

Languages: Python, C/C++, SQL, R, Git, Latex, JavaScript, Bash
ML/DL Frameworks: PyTorch, JAX, TensorFlow, Keras, Lightning, DeepSpeed, Megatron-LM, FairScale, Horovod, Ray, Apache Spark
LLM/GenAI Stack: LangChain, LangGraph, LlamaIndex, Semantic Kernel, AutoGen, CrewAI, DSPy, Guidance, vLLM, TGI, Triton
Model Development: Transformers, Diffusers, PEFT, LoRA, QLoRA, Accelerate, BitsAndBytes, Flash Attention, xFormers, Apex
Libraries: NumPy, Pandas, Scikit-learn, XGBoost, LightGBM, OpenCV, FAISS, Annoy, ONNX, TensorRT, Open- VINO, Gradio, Streamlit
MLOps & Monitoring: MLflow, Weights&Biases, Neptune, DVC, Kubeflow, Airflow, Prefect, BentoML, Seldon, ClearML
Databases: PostgreSQL, MongoDB, Redis, Elasticsearch, Qdrant, Pinecone, Weaviate, Milvus, LanceDB, ChromaDB **Cloud &**

Infrastructure: AWS (SageMaker, Bedrock, Lambda, ECS, EKS), GCP (Vertex AI, TPUs), Azure (ML Studio, OpenAI Service), CUDA Toolkit
Deployment: Docker, Kubernetes, FastAPI, gRPC, GraphQL, Terraform, GitHub Actions, ArgoCD, Jenkins, Prometheus, Grafana
Research Tools: Jupyter, Colab, Paperspace, Weights&Biases, Tensorboard, Optuna, Ray Tune, Hydra, ComfyUI

Education

Government College University Faisalabad — Faisalabad, Pakistan
Bachelor of Science in Data Science — *Oct 2021 – Jun 2025*

Institute of Emerging Careers — Lahore, Pakistan
Professional Certification in Data Analytics — *Jun 2022 – Jan 2023*

Certifications

Stanford University — Coursera
Machine Learning Specialization — *Issued 2023*

Google — Coursera
Google Data Analytics Professional Certificate — *Issued 2023*

PFTP — Pakistan
Certified Artificial Intelligence Professional — *Issued 2022*

Institute of Emerging Careers — Lahore, Pakistan
Data Analytics Certification — *Completed 2023*

Volunteer Experience

Google Developer Student Club (GDSC), GCUF Chapter — Faisalabad, Pakistan
Management Lead — *Sep 2023 – May 2024*

- Orchestrated **15+ technical workshops** and hackathons on AI/ML and cloud technologies for over **500+ stu- dent attendees**.
- Led a team of **10 core members** to plan event logistics, marketing, and speaker outreach, ensuring seamless execution.
- Grew student chapter membership by **150%** through strategic outreach campaigns and university-wide collabora- tions.