

M. Ikram Ullah

AI/ML Engineer · LLM Agents, RAG & Evaluation

Islamabad, Pakistan | (+92) 311-7862537 | ikramullah4748@gmail.com | linkedin.com/in/ikram-ullah11 | github.com/ikram4748 | ikram4748.github.io

SUMMARY

AI/ML engineer who builds LLM agents and retrieval systems and measures them. Recent work: a self-correcting text-to-SQL agent evaluated on held-out BIRD questions (+6.0 points execution accuracy over a one-shot baseline), a support agent that takes real actions behind a guardrail gateway (policy in code, human approval, idempotent refunds), a multi-agent literature reviewer with verified citations, and a production-style legal RAG assistant over 117,000 statute sections. Comfortable from data pipeline to FastAPI/Docker deployment; pursuing an MS in Data Science.

TECHNICAL SKILLS

LLM agents & RAG: LangGraph (state machines, interrupts, checkpoints), LangChain, function calling, multi-agent pipelines, hybrid retrieval (BM25 + dense, RRF), cross-encoder reranking, HyDE, Qdrant, FAISS, ChromaDB

Evaluation & safety: execution accuracy, RAGAS-style faithfulness, LLM-as-judge with validation, simulated-user (tau-bench style) evaluation, dev/test splits, paired bootstrap CIs, prompt-injection defences, guardrails

Models & APIs: Gemini, OpenAI and Anthropic APIs, Hugging Face Transformers, sentence-transformers, BGE-M3, ONNX Runtime / fastembed, Whisper ASR, open-weight LLMs (Qwen, Gemma, Llama)

ML & deep learning: PyTorch, TensorFlow/Keras, scikit-learn, XGBoost, LightGBM, CNNs, Vision Transformers, fine-tuning, SHAP

Engineering: Python, SQL (SQLite, DuckDB, PostgreSQL, MySQL), FastAPI, React, Docker, Git, pytest, Streamlit, n8n

EXPERIENCE

AI/ML Engineer, Syslab Technologies (Islamabad, Pakistan) Oct 2024 – Nov 2025

- Architected VESARA, an agentic healthcare voice and chat assistant combining a RAG pipeline (FAISS + Hugging Face sentence-transformers) with Whisper speech recognition and neural text-to-speech.
- Built n8n automation workflows orchestrating multi-step agent actions (appointment scheduling, symptom-triage routing, EHR lookup), improving response throughput by 31%.
- Improved diagnostic response accuracy by 23% through systematic A/B testing of model configurations, prompt strategies and RAG retrieval settings (chunk size, retrieval depth).
- Built an automated model-monitoring pipeline (accuracy, precision, recall, F1, latency) with real-time Power BI dashboards for early detection of model drift.
- Investigated model failures through error logs, edge cases and interaction data; targeted retraining and data-quality fixes reduced the error rate by 18%.

Freelance AI/ML Engineer, Self-employed (direct clients) (Remote) Feb 2023 – Present

- Delivered 15+ end-to-end AI/ML projects for healthcare, e-commerce and fintech clients, from scoping and data collection to deployment and handoff.
- Built RAG chatbots and document Q&A systems (LangChain, FAISS, Hugging Face embeddings) that answer from clients' internal knowledge bases with cited, grounded responses.
- Developed n8n workflows connecting LLM agents to CRM, e-mail and scheduling APIs, cutting clients' manual operational work by 30–45%.
- Fine-tuned Hugging Face models (BERT, DistilBERT, Mistral-7B) for text classification, sentiment analysis and named-entity recognition; deployed solutions with FastAPI and Docker.

PROJECTS

Customer-Ops Agent with Guardrails (github.com/ikram4748/customer-ops-agent)

LangGraph · Gemini function calling · SQLite · Streamlit

- Built a support agent that executes refunds, cancellations and address changes behind a gateway the model cannot bypass: identity and argument-provenance checks, policy enforced in code, customer confirmation codes, supervisor approval through LangGraph interrupts, and idempotency keys that stop double refunds when a payment times out.

- Designed a tau-bench-style evaluation of 45 simulated-customer scenarios (including prompt injection, social engineering and injected tool failures), scored on the final database state with automatic policy-violation detection, comparing a prompt-only agent with the guarded one.

Data Analyst Agent (text-to-SQL) (github.com/ikram4748/data-analyst-agent)

LangGraph · Gemini · sqlglot · DuckDB · fastembed · Streamlit

- Built a self-correcting text-to-SQL agent: local-embedding schema linking, a two-layer read-only SQL guard, a literal-value check that catches silent zero-result errors, and a bounded fix loop driven by database feedback.
- Raised execution accuracy on 150 held-out BIRD questions from 60.7% (one-shot baseline, same model) to 66.7% (+6.0 points, 95% CI +1.3 to +11.3) at 1.09 LLM calls per question; tuned on a separate dev split after error analysis of 48 failures.
- Found and fixed a measurement bug in which rate-limit fallback silently mixed three models across benchmark variants.

Literature Review Agents (github.com/ikram4748/research-assistant-agent)

LangGraph · Gemini · arXiv & OpenAlex APIs · fastembed

- Built a multi-agent reviewer (planner, searcher, reader, writer, critic, reviser) in which every finding must quote its source abstract word for word, checked in code, and a critic sends unsupported sentences back for revision.
- Evaluated citation grounding against a single-agent baseline with an independent LLM judge whose sensitivity is validated by swapping citations to wrong papers.

Financial RAG Lab (github.com/ikram4748/financial-rag-lab)

PyMuPDF · BM25 · bge-small · cross-encoders · Gemini

- Built a RAG benchmark on FinanceBench (150 analyst questions, 84 filings, 12,013 pages) comparing 3 chunking strategies, 3 metadata scopes and 5 retrievers (BM25, dense, hybrid RRF, two cross-encoder rerankers) plus HyDE, scored on evidence-page recall with a dev/test split.
- Measured end-to-end answer accuracy from closed-book to an evidence-page oracle, graded by a separate judge model validated on gold and corrupted answers.

Qanoon Assist: legal-information RAG for Pakistani law (private) (*code available on request*)

FastAPI · React · PostgreSQL · Qdrant · BGE-M3 · Gemini

- Scraped and parsed 6,664 federal and provincial laws (117,365 sections, 147,701 chunks) from five government portals; answers in English, Urdu and Roman Urdu with section-level citations.
- Hybrid BM25 + BGE-M3 dense retrieval with bge-reranker-v2-m3 and citation verification; hit@5 1.00, faithfulness 0.944 and 4.8 s average latency on a 36-question evaluation; 124 automated API and UI tests.

RESEARCH

UrduComicLab: complete-story Urdu story-to-comic generation (MS thesis, in progress)

- Open-weight pipeline turning complete Urdu stories into right-to-left comics; vs baseline: spoken lines printed 0.49 → 1.00, dialogue F1 0.63 → 1.00, face consistency 0.54 → 0.67 (paired bootstrap).

Facial emotion recognition on imbalanced FER2013: CNN vs Vision Transformer (paper, pending publication)

- CNN (SE blocks) vs ViT-Base/16 on FER2013 with class-balanced focal loss; ViT avoids minority-class collapse (Disgust 64.9% vs 0%).

EDUCATION

MS Data Science, Quaid-e-Azam University, Islamabad

Feb 2025 – Present

Coursework: Machine Learning, Deep Learning, NLP, Big Data Analytics, Statistical Modeling, Data Visualization. Thesis: UrduComicLab.

CERTIFICATIONS

Machine Learning Specialization · Deep Learning Specialization · Natural Language Processing · Big Data Analytics