

M. Ikram Ullah

ML Researcher · Multimodal Generation, LLM Evaluation & Applied Deep Learning

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

SUMMARY

Machine-learning researcher (MS Data Science, Quaid-e-Azam University) working on low-resource multimodal generation and on measuring LLM systems rigorously. My MS thesis turns complete Urdu stories into right-to-left comics with open-weight models, with paired-bootstrap comparisons against a baseline. Earlier work compares CNNs and Vision Transformers on imbalanced facial-emotion data. I design evaluations with held-out splits, ablations, confidence intervals and validated LLM judges.

RESEARCH

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

- Turns complete Urdu children's stories into right-to-left comic books with correctly shaped Nastaliq dialogue, narration captions and sound effects, using open-weight models only (Gemma-2-9B parser, SDXL with IP-Adapter, Grounding DINO bubble placement, HarfBuzz lettering).
- Against a baseline on 15 paired comics (3 stories × 5 seeds, paired bootstrap CIs): spoken lines printed 0.49 → 1.00, dialogue F1 0.63 → 1.00, speaker accuracy 0.83 → 1.00, face consistency (DINOv2) 0.54 → 0.67; about 35 pinned open models benchmarked across four experiment campaigns.

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

- Compared an enhanced ResEmoteNet CNN with squeeze-and-excitation blocks against ViT-Base/16 on FER2013 (35,887 images, 7 classes, 16.5:1 class ratio), with class-balanced focal loss, two-stage training and CutMix, Mixup and RandAugment.
- Showed that ViT self-attention avoids minority-class collapse (64.9% vs 0% accuracy on Disgust), while the CNN with channel attention was stronger overall on low-resolution faces.

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.

PROJECTS

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.

E-mail Experiment & Uplift Modeling (github.com/ikram4748/uplift-ab-testing)

Python · statsmodels · LightGBM · Plotly · Streamlit

- Analysed a 64,000-customer randomized e-mail experiment end to end: sample-ratio and balance checks, Holm-corrected effects (men's e-mail +7.7 points visits, +\$0.77 spend per customer, 95% CI \$0.49–\$1.05), CUPED and regression adjustment, power analysis, and a peeking simulation (10 looks → 18% false positives).
- Built S/T/X-learner and transformed-outcome uplift models with Qini evaluation over 5 splits and valued targeting policies by inverse-propensity weighting; showed that e-mailing everyone beats model targeting below about \$0.23 per e-mail.

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.
- Improved diagnostic response accuracy by 23% through systematic A/B testing of model configurations, prompt strategies and RAG retrieval settings (chunk size, retrieval depth).
- 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.
- Fine-tuned Hugging Face models (BERT, DistilBERT, Mistral-7B) for text classification, sentiment analysis and named-entity recognition; deployed solutions with FastAPI and Docker.

TECHNICAL SKILLS

Research methods: experimental design, ablations, held-out dev/test protocols, paired bootstrap CIs, error analysis, LLM-as-judge validation, reproducible pipelines (pinned model revisions, fixed seeds)

Deep learning: PyTorch, Vision Transformers, CNNs (SE blocks), focal and class-balanced losses, CutMix/Mixup, diffusion models (SDXL, FLUX), IP-Adapter, Grounding DINO, DINOv2

NLP & LLMs: open-weight LLMs (Gemma, Qwen, Llama), Urdu/Nastaliq NLP and RTL text shaping (HarfBuzz), RAG, LLM agents (LangGraph), Hugging Face Transformers

Statistics & data: causal inference, A/B testing, uplift modeling, Python, SQL, pandas, statsmodels, scikit-learn, LightGBM

Engineering: GPU experiment campaigns, Git, Docker, FastAPI, pytest

CERTIFICATIONS

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