Syed Moiz Ali

+92-324-4681684 | moizsyedali.ma@gmail.com | linkedin.com/in/moizali01 | github.com/moizali01

Education

Lahore University of Management Sciences Lahore, Pakistan
Bachelor of Science in Computer Science *Sep. 2021 – May 2025*

Relevant Coursework: Topics in Computer and Network Security, Deep Learning, Machine Learning, Network Security, Topics in Large Language Models, Computer Vision, Probability, Algorithms

Research Experience

LLM-Integrated Source Code Debloating Mar 2024 – Present
Security & Privacy Lab, LUMS *Lahore, Pakistan*

- **Collaborators:** Dr. Fareed Zaffar (LUMS), Dr. Ashish Gehani (Stanford Research International), Dr. Sazzadur Rahman (University of Arizona), Dr. Fahad Shaon (Google).
- Manually debloated five programs to establish ground truth benchmarks for auditing and evaluating source code debloating tools.
- Leveraged **LLVM coverage** and code semantics to introduce stability-focused heuristics, reducing critical functionality loss while maintaining program security.
- Developed a novel **RAG**-based multiagent LLM pipeline to assist traditional debloaters by retaining code critical for functionality and generality in debloated software.
- Designed specialized LLM prompts informed by manual analysis to address limitations in current debloaters, enhancing the relevance of retained code.
- Achieved improved generality and stability across benchmarks when integrated with three existing debloaters, with minimal size impact.

Generalizability of Knowledge Injection in Language Models Nov 2024 – Present
Security & Privacy Lab, LUMS *Lahore, Pakistan*

- **Collaborators:** Dr. Fareed Zaffar (LUMS), Dr. Yasir Zaki (NYU Abu Dhabi)
- Building on our previous research that identified limitations in the generalization of LLM safety across tasks, this project examines the generalizability of post pre-train knowledge injection.
- Currently investigating how knowledge injected in LLMs transfers and generalizes across a variety of tasks.

MultitaskBench: Cross Task LLM Safety Alignment Feb 2024 – Sep 2024
Security & Privacy Lab, LUMS *Lahore, Pakistan*

- **Collaborators:** Dr. Fareed Zaffar (LUMS), Dr. Yasir Zaki (NYU Abu Dhabi), Faizan Ahmad (Security Engineer, Meta).
- Investigated task-specific safety degradation in finetuning of LLMs, uncovering vulnerabilities in tasks such as code generation, translation, and classification.
- Conducted a comprehensive analysis of existing safety solutions, highlighting limitations in fine-tuning datasets, external guard mechanisms, and model alignment techniques.
- Developed and curated MultitaskBench a safety alignment dataset that enhances safety across various LLM tasks. Published at COLING 2025

LLM Hallucination Benchmarking Oct 2024 – Present
LUMS *Lahore, Pakistan*

- Designed a probabilistic framework to detect hallucinations, analyzing log probabilities and top-k token distributions.
- Conducted initial benchmarks to identify patterns distinguishing confident and hallucinated responses.
- Explored threshold-based metrics to improve error identification, with preliminary findings guiding further refinement of the methodology.

Publications

Multitask Mayhem: Unveiling and Mitigating Safety Gaps in LLMs Fine-tuning | *arXiv:2409.15361* 2024

- Essa Jan, Nouar AlDahoul, **Moiz Ali**, Faizan Ahmad, Fareed Zaffar, Yasir Zaki.
- Investigates task-specific safety gaps in fine-tuned LLMs and proposes a multitask safety dataset to mitigate them.
- Published at **COLING 2025**.

Experience

Teaching Assistant Sep 2023 – Present
Lahore University of Management Sciences *Lahore, Pakistan*

- Designed and evaluated programming assignments and quizzes for a graduate-level course on **Computer Vision Fundamentals** (Fall 2024, Dr. Murtaza Taj).
- Conducted tutorials, graded assignments for 90 students, and provided individual support for **CS100: Computational Problem Solving** (Spring 2023, Dr. Fareed Zaffar).
- Led tutorials, graded labs and quizzes, and provided individual assistance for **CS200: Introduction to Programming** (Fall 2023, Dr. Shafay Shamail).

Projects

Tradesnap.ai | *MERN, Selenium, Azure Cloud, OpenAI* Jan 2024 – May 2024

- Developed a conversational stock trading platform using OpenAI's Assistant to enable multilingual stock trading via chat interface.
- Integrated features like buying/selling stocks, educational content, and personalized volatility alerts.
- Scraped data from PSX for platform backend and built detailed company pages with advanced React charts.
- Implemented automated testing for the application using Selenium to ensure platform reliability.

Nighttime Wildlife Monitoring | *CycleGAN, Image Processing, OpenAI CLIP* Jan 2024 – May 2024

- Developed a hierarchical model leveraging CycleGANs to enhance nighttime camera trap images for snow leopard detection.
- Used OpenAI's CLIP for image classification and fine-tuned it for challenging nighttime conditions.
- Collected and curated training data from the Snapshot Serengeti Database, achieving 0.95 accuracy and 0.89 F1-score.

Urban Electricity Analytics | *Selenium, LSTM, Python, Pandas* Jun 2023 – Aug 2023

- Developed a high-performance web scraper using **Selenium** and multithreading to extract electricity consumption data for over 3 million users across Lahore.
- Engineered an **LSTM**-based time series forecasting model to predict feeder overloading, improving grid management strategies.
- Conducted analysis of seasonal consumption patterns to identify **poverty hotspots**.

Social Media Toxicity Classifier | *Llama2, PEFT, Jigsaw Dataset* Jan 2024 – May 2024

- Developed a model to detect and flag harmful social media content, fine-tuning Llama2-7B (PEFT) for toxicity classification.
- Achieved 90% accuracy and an F1-score of 0.89 across 6 toxic classes using the Jigsaw Toxic Comment Classification Dataset.
- Reached a ROC of 0.85, ensuring effective detection of harmful content.

Technical Skills

Languages: Python, JavaScript, C, C++, Haskell, HTML, CSS, Bash
Technologies/Frameworks: PyTorch, TensorFlow, OpenCV, MERN, TypeScript, LLVM, LangChain, Pandas, Scikit-learn, LlamaIndex, OpenAI Platform, Google AI Studio, Selenium, Azure Cloud