# Muhammad Ayain Fida Rana

Cambridge, UK | +44 7456 672518 | mafr2@cam.ac.uk | /in/m-ayain-fida-rana | ayainfida.github.io | ayainfida

## Education

**University of Cambridge** — Cambridge, UK
*Master of Philosophy in Advanced Computer Science - MPhil ACS* — Oct 2025 – Jun 2026

- **Relevant Coursework:** Machine Visual Perception, Machine Learning and the Physical World, Mobile Health, Computer Security, Cryptography and Protocol Engineering

**Lahore University of Management Sciences (LUMS)** — Lahore, Pakistan
*Bachelor of Science in Computer Science - BS CS* — Sep 2021 – Jun 2025

- **CGPA:** 3.97/4.00 (Graduated with High Distinction)
- **Relevant Coursework:** Artificial Intelligence, Computer Networks, Data Science, Distributed Systems, LLM Systems, Machine Learning, Network Security, Operating Systems, Software Engineering

## Publications

**Semantic Caching for Improving Web Affordability**

Hafsa Akbar, Danish Athar, **Muhammad Ayain Fida Rana**, Zartash Afzal Uzmi, Ihsan Ayyub Qazi, Zafar Ayyub Qazi ***(Under review at WWW 2026)***

**LLM-Enabled Semantic Caching For Affordable Access**

Hafsa Akbar, Danish Athar, **Muhammad Ayain Fida Rana**, Zartash Afzal Uzmi, Ihsan Ayyub Qazi, Zafar Ayyub Qazi ***(Poster, WiML @ NeurIPS 2025)***

## Research Experience

**Distributed and AI Systems Lab, LUMS** — Lahore, Pakistan
Research Assistant — Jun 2024 – May 2025

- Reduced bandwidth costs for online news and media platforms by ∼**10%** through a semantic caching system that reused similar images across articles, cutting total page weight by up to **6.4%.**
- Built an automated selenium-based scraping pipeline to collect and process **4,264** images with metadata (headlines, alt text) from **50** leading global news platforms for contextual analysis.
- Annotated **40,000+** image pairs for similarity analysis, identifying high potential categories (gender, business, sports) with up to **37%** reusability, and released a public dataset for the research community.
- Engineered a two-step LLM pipeline (LLaVA-NeXT + LLaMA 3.1) that achieved **91%** precision and **63%** recall, delivering a reliable (no page-breaking) yet conservative performance comparable to commercial multi-modal models in assessing replaceability.
- Research findings accepted at WiML @ **NeurIPS 2025**.

**Course Project: Approximate Caching for Fact Checking** — Lahore, Pakistan
Supervised by Dr. Zafar Ayyub Qazi and Dr. Ihsan Ayyub Qazi, LUMS — Sep 2024 – Dec 2024

- Designed a multilingual approximate caching system using OpenAI's text-embedding-3-large and FAISS to detect recurring fact-checking claims.
- Calibrated similarity thresholds to **0.8**, achieving **96.1%** agreement with human verification and **99.8%** agreement with ground-truth verdicts.
- Analyzed temporal and linguistic trends, revealing short-term misinformation recurrence and lower cache reuse for localized languages, and efficiency gains for global fact-checking organizations by reusing verified claims and reducing verification latency.

**Networks and Systems Group, LUMS** — Lahore, Pakistan
Research Assistant — Aug 2023 – May 2024

- Motivated by **SIGCOMM'23** findings that cache savings dropped from **60.9%** to **21.4%** due to device memory limits, investigated mobile caching behaviors to uncover inefficiencies and inform optimization strategies.
- Identified critical gap in Chrome mobile caching documentation, conducted cache measurement experiments that revealed cache expansion to nearly **100%** of device storage before eviction, regardless of the memory limits, and motivating to intelligently reuse content (semantic caching) beyond traditional (exact) caching.
- Automated large-scale performance testing across **10,000+** websites using Appium, DevTools, and ADB Shell to collect and process cache contents/headers, storage utilization, and memory usage for analyzing cache eviction policies.

**Networks and Systems Group, LUMS** — Lahore, Pakistan
Research Intern — May 2023 – Aug 2023

- Designed and conducted a user study with **35** participants to benchmark a **SIGCOMM'23** framework against Brave and Opera Mini, delivering **11%** and **7%** greater page weight reductions respectively and achieving **50%** higher user satisfaction scores.

## Teaching Experience

**CS 582: Distributed Systems** — Lahore, Pakistan
Teaching Assistant — Sep 2024 – Dec 2024

- Conducted weekly office hours and tutorials for over 70 students, created and graded quizzes, and implemented automated grading for assignments.
- Managed the course Slack channel, addressing student queries and facilitating discussions to enhance learning.

**CS 310: Algorithms** — Lahore, Pakistan
Teaching Assistant — Sep 2024 – Dec 2024

- Supported students on course's Slack channel, and engaged in semi-formal student counseling.
- Conducted weekly office hours for over 200 students, created/invigilated/graded quizzes, and provided feedback on homeworks.

**CS 202: Data Structures** — Lahore, Pakistan
Teaching Assistant — Jan 2024 – May 2024

- Managed course's Slack channel, created/reviewed/invigilated/graded quizzes and programming assignments.
- Held weekly office hours for over 100 students, providing additional academic support and guidance to students.

## Awards & Honors

- Awarded the **Vicky Noon Scholarship** (Cambridge Trust) for **2025–26**.
- Graduated with **High Distinction**, ranked in the **top 3%** of the LUMS SBASSE Class of **2025**.
- Placed on Dean's Honor List for **2021-22, 2022-23, 2023-24, 2024-25**.
- Awarded Merit Scholarship (LUMS) for **2022-23, 2023-24, 2024-25**.
- **Top in World** in A Level Mathematics in **2020**.
- Roll of Honor **(Highest Student Award)** at Beaconhouse Johar Town in **2019**.

## Development Projects

**Succession Planning Portal** | *React, JavaScript, Node.js, MongoDB, TensorFlow*

- Built a full-stack HR portal prototype for centralized tracking of performance, skills, and feedback using **dummy** employee data to simulate promotion-readiness assessments.
- Implemented regression models in TensorFlow and integrated them into a scalable React/Node.js HR portal, generating promotion predictions with career path visualizations to reduce subjective bias and validate feasibility for enterprise deployment.

**The Bean Journal** | *Next.js, React, Node.js, MongoDB, Google Maps API, FreeImage API*

- Developed a full-stack coffee review platform with interactive maps, photo uploads, and role-based access control to ensure authentic and discoverable reviews.
- Deployed on Vercel with MongoDB Atlas, hosting reviews from **4** countries with scalable search and seamless performance.

**Sarmaya Car: Intelligent Used Car Recommender** | *Python, Selenium, Pandas, PuLP*

- Collected and processed **66,000+** car listings from PakWheels.com, enabling scalable depreciation and trend analysis across 8 years of market data to better inform buyer decision-making.
- Developed a first-of-its-kind goal-programming recommender for optimal car selection under user-defined priorities, cutting average buyer decision time from **30** minutes to under **10**, which led to an invitation from PakWheels to explore deployment.

**Distributed, Fault-Tolerant Key-Value Store** | *Go*

- Implemented a key-value store on top of the Raft consensus algorithm, based on the paper "*In Search of an Understandable Consensus Algorithm*", and demonstrated strong consistency and availability across a **5**-node cluster.
- Handled client operations (Get, Put, Append) with deduplication for exactly-once semantics, supporting concurrent requests safely under leader changes and network partitions.
- Validated fault tolerance through automated tests simulating leader crashes, partitions, restarts, and **10,000+** client operations, demonstrating reliable recovery and agreement across replicas.

**SastaGPT** | *Python, PyTorch, NumPy, Matplotlib, Pandas*

- Implemented a Transformer model from scratch in PyTorch with embeddings, multi-head attention, and positional encodings, based on the paper *"Attention Is All You Need"*.
- Trained on a **100k+** token dataset using subword tokenization (GPT-2 encoder), optimizing training stability with GELU activations and dropout.
- Achieved stable convergence and generated coherent, character-specific text sequences, demonstrating the architecture's effectiveness compared to baseline RNNs in producing contextually consistent outputs.

**RAG-Based Researcher Chatbot** | *Python, LangChain, Pinecone, FAISS*

- Built a Retrieval-Augmented Generation (RAG) chatbot using LangChain, FAISS, and Pinecone, integrating **10** research papers with Wikipedia for source-cited responses.
- Improved answer reliability by applying citation-grounding and custom prompt templates, which reduced hallucinations in evaluation queries by an observed ∼**70%**; hence, enabling accurate, verifiable responses for research assistance use cases.

**LLM-Powered Evaluation System** | *Python, Regex, LaTeX, Pandas*

- Developed an automated assignment grading system combining regex-based extraction from LaTeX files with LLM grading using few-shot prompting and chain-of-thought (CoT) reasoning.
- Achieved **96%** grading accuracy against instructor rubrics while maintaining higher consistency than human graders (**85%** inter-rater reliability), and enabled near-instant feedback that only required a final manual review pass by the grader.

**Command Line Shell** | *C*

- Programmed a minimal command-line interpreter that emulates core UNIX shell functionalities, including support for I/O redirection, piping output between commands, wildcards, and chaining commands in sequence.

**User Level Threading Library** | *C*

- Created a fairly abstracted threading library that, although utilized registers for storing PCBs, did application-level context switching.
- Implemented a Round Robin scheduler for thread management and developed concurrency and synchronization primitives to handle thread coordination and avoid conflicts.

**Simple File System** | *C*

- Developed a UNIX-like file system with partitions for superblocks, inodes, and datablocks, supporting file reading and writing, and operating between a simple shell program and a disk emulator.

## Skills

**Languages:** Python, JavaScript/TypeScript, C/C++, SQL, Go, Bash, MATLAB, Haskell, VBA
**Frameworks:** React, Node.js/Express, PyTorch, TensorFlow, scikit-learn, Pandas, NumPy, Keras, Flask, FastAPI, OpenCV, Selenium
**Cloud/Tools:** AWS, ADB, DevTools, Android Studio, GCP, Docker, Git/GitHub, Linux, MongoDB, Redis, Jupyter, Postman, VS Code