

MUHAMMAD ASSAD ULLAH

On-Device AI Systems | Mobile Software Engineering | Edge Intelligence

asadullahai602@gmail.com | +92 311 602 3374 | Hasilpur, Punjab, Pakistan | Passport: DA1873951

[GitHub](#) | [LinkedIn](#) | [Portfolio](#) | [ORCID: 0009-0007-2290-304X](#)

RESEARCH PROFILE

Final-year Computer Science undergraduate at UET Taxila, Pakistan, currently working remotely as an **Edge AI Research Engineer at Duke Kunshan University's Edge Intelligence Lab** under Associate Professor Bing Luo. My work is in on-device LLM inference, mobile systems engineering, and edge-AI reliability. I write native C/C++ bindings, manage tensor memory by hand, and build fault-injection pipelines to find what breaks when hardware degrades. I do not just train models; I get them running reliably on consumer phones. My open-source work has been adopted by the FSE group at East China Normal University. Expected graduation: July 2027.

RESEARCH AND PROFESSIONAL EXPERIENCE

Edge AI Research Engineer (Remote)

[Duke Kunshan University, Edge Intelligence Lab](#) | Kunshan, China (Remote)

Aug 2026 to Present | Supervisor: [Prof. Bing Luo, Associate Professor](#)

- Built a fully on-device Android fine-tuning pipeline for **Qwen3.5-2B VLM** using **PEFT/LoRA**; the complete training loop runs on consumer Android hardware with no cloud dependency.
- Engineered **JNI and dart:ffi native bindings** into **llama.cpp** and **libmtmd**, giving Flutter apps direct access to on-device multimodal inference with manual tensor memory management. Configured **Vulkan GPU acceleration** via custom llama.cpp bridges.
- Integrated the SATE AI fault-injection framework into the lab evaluation pipeline to measure model resilience under memory pressure and thermal throttling.

Trainee Artificial Intelligence Engineer

DevCrew I/O | Lahore, Pakistan

Jul 2026 to Aug 2026

- Built production RAG pipelines connecting LLMs with vector databases and REST APIs.
- Trained and evaluated ML models; handled dataset preprocessing and benchmarking.

Trainee Software Development Engineer

DevCrew I/O | Lahore, Pakistan

Jun 2025 to Aug 2025

- Completed a full mobile development lifecycle in Flutter; shipped production features across client apps.
- Integrated RESTful APIs, wrote unit tests, and worked in an Agile team alongside senior developers.

EDUCATION

Bachelor of Science in Computer Science

[University of Engineering and Technology \(UET\) Taxila](#) | Pakistan

Aug 2023 to Jul 2027 (Expected) | CGPA: 3.31 / 4.00 | EQF Level 6

Thesis: AURAF (Adaptive Unified Runtime Architecture for Flutter) -- unifying token budgeting, cloud/edge fallback, and hardware-aware on-device AI for Flutter apps

Coursework: Artificial Intelligence, Data Structures and Algorithms, Operating Systems, Database Systems, Computer Networks, NLP, Compilers, Software Engineering

Intermediate in Computer Science (ICS)

[Punjab Group of Colleges](#) | DG Khan, Pakistan

Aug 2021 to May 2023 | Score: 1001/1100 | 3rd Position, D.G. Khan Board (2023)

PUBLICATIONS AND MANUSCRIPTS

[1] Assad Ullah, M. (2026). *Characterising On-Device AI Failure Modes Under Joint Fault Injection and Runtime Observability*. Zenodo. [doi:10.5281/zenodo.23262270](#)

[2] Assad Ullah, M. (2026). *SATE AI: A Fault Injection and Reliability Engineering Framework for On-Device AI Models*. Zenodo. [doi:10.5281/zenodo.23250418](#)

[3] Assad Ullah, M. (2026). *EdgePulse: A Runtime Observability Framework for Quantized LLMs on Consumer Edge Devices*. Zenodo. [doi:10.5281/zenodo.23248718](#)

SATE AI paper draft prepared for submission to JOSS and arXiv.

RESEARCH PROJECTS AND OPEN-SOURCE CONTRIBUTIONS

SATE AI: Fault-Injection Framework for On-Device AI [pub.dev](#) | [GitHub](#) | [Zenodo Paper](#) | MIT | 9 releases

- Stress-tests on-device AI models under memory pressure, corrupted input, thermal throttling, and latency spikes.
- Ships 11 fault injectors, 8 model adapters (ONNX, TFLite, llama.cpp, MediaPipe), and a 246-case automated test matrix. pub.dev quality score 160/160.

EdgePulse: Runtime Observability SDK for Edge LLMs [pub.dev](#) | [GitHub](#) | [Zenodo Paper](#)

- Real-time telemetry capturing memory RSS, thermal state, battery draw, and CPU utilization per inference trace on Android and iOS.
- Found a thermal-API blind spot in the Android OS: the standard API fails to report CPU frequency throttling under heavy LLM workloads.

Offline Multimodal Vision Assistant [GitHub](#) | Flutter | llama.cpp | Qwen3-VL-2B

- Fully local on-device Android app running Qwen3-VL-2B offline. Full tokenization and inference pipeline within mobile memory limits.
- Configured Vulkan GPU acceleration via custom llama.cpp bridges; tuned memory layout for dense visual patch embeddings.

u2_flutter: Flutter UI Testing Driver [GitHub](#) | [uiautomator2 PR #1196](#) | PyPI

- Extended uiautomator2 (7,800+ GitHub stars) to support Flutter apps; officially listed as a supported plugin in the [repository](#).
- **Adopted by the FSE group at East China Normal University:** merged into their [Kea2 framework](#) in PRs #229-234 and flagged as a starting point for a forthcoming FSE publication.

AndroTest24 Extension [ASE 2024 artifact repository](#)

- Added Flutter support to the AndroTest24 GUI testing benchmark suite. PR #2 merged; PR #4 under review.

AI Student Assistant [GitHub](#) | [Live App](#) | Flutter | Firebase | OpenRouter

- Multi-key API rotation via OpenRouter handles 200 daily requests with automated failover tracking.
- Offline-first: SQLite local caching paired with real-time Firebase Firestore sync.

PUBLISHED PACKAGES (PUB.DEV)

[sate_ai](#) | [edgepulse](#) | [edgepulse_core](#) | [edgepulse_cli](#) | [offline_sync_flutter](#) | [flutter_form_guard](#)

TECHNICAL SKILLS

Languages: Dart/Flutter, Python, C/C++, JavaScript, SQL

On-Device AI: llama.cpp, GGUF models, Qwen-family VLMs, ONNX Runtime, PEFT/LoRA, PyTorch, HuggingFace, TFLite

Mobile Systems: Android JNI, dart:ffi, native bindings, Vulkan GPU, Riverpod, Hive, Melos, Firebase/Firestore, MVVM

Testing and Ops: Fault injection, property-based UI testing, RAG pipelines, CI/CD, Git/GitHub, pub.dev

Data and Analytics: PySpark, T-SQL, Power BI, Tableau, SQLite

HONOURS, AWARDS AND ACHIEVEMENTS

1st Place, Techinance CyberHack 2025 2025 | Judged by representatives from Amazon and the University of Alabama.

7x Hackathon Winner 2023 to 2026 | Wins across AI, Flutter, cybersecurity, and sustainability competitions.

Official Plugin Contributor, uiautomator2 (7,800+ GitHub stars) 2025 | u2_flutter officially listed as a supported plugin. PR #1196.

3rd Position, D.G. Khan Board, Intermediate in Computer Science 2023

Graduate, Aspire Leaders Program, Harvard Graduate School of Education 2024

CERTIFICATIONS

Developing Mobile Apps with Flutter Specialization | IBM / Coursera Dec 2025

Google Prompting Essentials Specialization | Google / Coursera Jun 2025

Artificial Intelligence Using Python | DigiSkills / Virtual University of Pakistan Jul 2026

Master Flutter and Firebase | Udemy

Dart and Flutter: Complete Development Course | Udemy

LEADERSHIP AND EXTRACURRICULAR

Finance Secretary

Umeed-e-Subh Welfare Society (Student NGO), UET Taxila

2024 to Present

- Managed annual operating budgets and maintained ledger tracking for all community-driven initiatives.
- Coordinated corporate fundraising and sponsorship drives; prepared end-of-quarter financial audits.

Executive, Marketing

Microsoft Learn Student Ambassadors, UET Taxila

2023 to 2026

- Scaled attendee registration for developer hackathons and code challenges through targeted outreach.
- Promoted cloud computing workshops and Microsoft developer ecosystems across campus.

LANGUAGES

Urdu: Native | Punjabi: Native | English: Professional Working Proficiency (B2)

REFERENCES

Prof. Bing Luo | Current Research Supervisor

Associate Professor, Edge Intelligence Lab, Duke Kunshan University, Kunshan, China

bing.luo@dukekunshan.edu.cn

Dr. Aun Irtaza | Academic Supervisor

Chairman / HOD, Department of Computer Science, UET Taxila, Pakistan

aun.irtaza@uettaxila.edu.pk