

Abdur Rafay

abdurrafay8709@gmail.com | LinkedIn | GitHub | Portfolio

Research Interests

Agentic AI systems for software engineering: retrieval-augmented generation, automated program repair, multi-agent orchestration, and the reliability and security of LLM-integrated developer tools.

Education

FAST NUCES – BS in Computer Science, GPA: 3.51/4.00

Aug 2025 – Present

- **Awards:** Dean's List (Fall 2025); 1st Place, Academic Merit Scholarship (KIPS Prep. Division); 157th Nationally, NSTC Mathematics Olympiad (NMTC-21)
- **Certifications & Activities:** Generative AI Development (ULEFUSA), Full Stack Web Development (ICT Trainings); ICAP Math Championship, HEC GenAI Hackathon, SOFTEC App Dev Hackathon

Projects

Patchwork – Automated Program Repair Agent | *Python, LangGraph, Ollama, Pydantic, Pytest, Ruff* [Github]

- Architected an offline program repair agent (LangGraph, 4-bit Qwen2.5-Coder-3B), reducing per-defect repair latency by 74% (29.6s to 7.8s) under 2.5GB VRAM.
- Boosted defect repair rates from 38% Pass@1 to 58% Pass@overall across a 100-defect benchmark via an execution-guided reflection harness (AST parsing, Ruff linting, sandboxed pytest execution).
- Discovered a 67% confirmation-bias failure mode across 21 fault domains where generated test suites passed on fixed code but failed oracle tests, isolating systematic reasoning blind spots in small code LLMs.

InsafDost AI – Asynchronous Legal RAG Pipeline | *Python, FastAPI, LangGraph, Qdrant, Docker* [Github]

- Built an asynchronous LangGraph pipeline in FastAPI, dynamically routing between Groq-hosted 20B and 120B parameter models to stream appellate legal analyses via SSE.
- Improved retrieval accuracy by 19.0% relative on Hit@1 (31.5% to 37.5%) and increased MRR@3 by 11.7% (0.383 to 0.428) across 200 queries by pairing Qdrant vector retrieval with a thread-offloaded BAAI cross-encoder reranker operating at a 0.40 sigmoid threshold.
- Achieved a 0.990 F1-score (100% adversarial rejection) on safety guardrails, and a 0.780 discrimination gap between grounded and fabricated authorities.

MediCore – Custom C++17 Clinical Application | *C++17, SFML, CMake, OOP* [Github]

- Developed a zero-dependency C++17 clinical system featuring custom dynamic array templates with geometric capacity doubling and schema-specific CSV serialization.
- Built an event-driven GUI directly on raw SFML 2D primitives, designing custom reusable UI component abstractions (input fields, tables, buttons) driven by an isolated rendering loop.

Experience

Flutter App Development Intern, The Xpertz – Lahore, Punjab, Pakistan

June 2024 – Aug 2024

- Implemented real-time client-side data synchronization in Flutter using Firebase Firestore listeners and REST endpoints, enabling dynamic record updates across client sessions.
- Stabilized a production mobile application by resolving 15+ UI and state defects, modularizing shared UI widgets, and implementing automated unit tests for API parsing models.

Skills

Languages: Dart, Python, TypeScript, C++, SQL, JavaScript, x86 Assembly

Frameworks & Libraries: Flutter, FastAPI

AI & Agentic Systems: LangGraph, Multi-Agent Orchestration, Small Language Models (SLMs), RAG, Semantic Vector Search (Qdrant), Cross-Encoder Reranking, Execution-Guided Reflection

Tools & Practices: Git, GitHub Actions (CI/CD), Docker, Firebase, Linux/Unix, Pytest, Ruff, AST Parsing, Ollama, RESTful APIs, Server-Sent Events (SSE)