

Muhammad Subhan Amir

03104072743 | subhanamir102@gmail.com | [linkedin.com/in/subhan-amir-8b4a32295](https://www.linkedin.com/in/subhan-amir-8b4a32295) | github.com/subhan-17h

CS undergraduate building production AI systems — shipped a Google Play app, an agentic NL2SQL engine, and edge-deployed computer-vision models. Focus: LLM agents, RAG, and evaluation-driven development.

EDUCATION

Information Technology University of the Punjab

Bachelor of Science in Computer Science

Lahore, Pakistan

2023 – 2027

EXPERIENCE

AI Engineer Intern

Qult Technologies

May 2026 – Jun 2026

Hybrid (Onsite + Remote), Lahore

- Built the core of an agentic **NL2SQL analytics engine** (UniBot) answering natural-language questions over institutional academic data, replacing a deterministic runtime and lifting benchmark accuracy to **50/50 (100%)** from an ~85% single-view / 0% cross-schema baseline.
- Engineered a **multi-node LLM pipeline** — Selector/router, Planner, and a bounded Think-Act-Observe loop of SQL generation, validation, execution, and refinement — with **tiered model routing** per node for cost control and cross-schema reasoning via sequential per-template SQL.
- Implemented a hard **safety-guardrail layer** (read-only DB role, single-SELECT validation, allowlisted schema, statement timeouts) proven by a test suite that blocks DDL/DML/injection/multi-statement SQL from reaching the database.
- Delivered a precision-first **semantic cache** and a **React chat UI** with a live agent trace, cutting answered-query latency from 13–18s to under 3s.

PROJECTS

Qemat — **Price Comparison Platform** | *Flutter, Dart, Firebase, Next.js, FastAPI, PostgreSQL*

- Designed and shipped a multi-client production system: a Flutter mobile app (released on Google Play Store) and a React/Next.js **web app** over a shared FastAPI backend for search, product matching, favorites, auth, and analytics.
- Built a hybrid data architecture where FastAPI serves product/trending/match APIs, verifies Firebase tokens (email/password + Google), and supports server-side filtering, sorting, and pagination over PostgreSQL-ingested product bundles.
- Implemented an offline-first delivery flow with bundle metadata/version checks, local cache persistence, and split download/async parse to cut cold-start latency with graceful fallback on network failure.
- Engineered an automated price-update pipeline (Firestore export → multi-store scraping → change validation against price history → batched Firestore updates) deployed to cloud with weekly runs.

Pothole Detection & Segmentation for ADAS | *Python, PyTorch, Ultralytics YOLO11, OpenCV, Core ML*

- Developed an ADAS road-hazard perception pipeline by fine-tuning YOLO11n for real-time pothole detection and YOLO11n/s-seg for instance segmentation on custom YOLO-format road-damage datasets.
- Achieved **0.9821 mAP@50** / **0.8473 mAP@50:95** for detection and up to 0.9856 box / 0.9816 mask mAP@50 for segmentation on held-out test data, with reproducible train/val/test workflows and metric export.
- Built OpenCV driving-video inference pipelines and validated on-device deployment by exporting to Core ML and running on mobile, enabling future driver-warning and path-planning use cases.

MYMIND — **Personal Academic Assistant** | *React, FastAPI, PostgreSQL, OpenAI Embeddings, Groq LLM*

- Built a full-stack student productivity platform (Next.js + TypeScript frontend, FastAPI backend) exposing modular REST APIs for academic planning and assistant interactions.
- Engineered a deterministic context layer (urgency, deadlines, missed items, free slots, workload warnings) feeding a hybrid daily planner that applies rule-based prioritization first, then LLM refinement.
- Implemented a hybrid RAG chat with query routing, course-scoped retrieval, **text-embedding-3-small** embeddings, and pgvector semantic search over chunked documents; integrated Google Classroom/Drive via OAuth 2.0.

PMPEDIA — **PM Standards Agentic RAG** | *React, FastAPI, ChromaDB, sentence-transformers, OpenAI API*

- Built a full-stack agentic RAG system over PMBOK7, PRINCE2, and ISO 21502 with domain-guarding, intent classification, per-standard query generation, coverage scoring, and corrective query rewriting before answer synthesis.
- Engineered semantic retrieval over locally ingested/chunked standards using sentence-transformer embeddings and persistent ChromaDB, with metadata-rich citations (section IDs, deep links, relevance scores) and agent-trace observability.

MindDrop — Privacy-First Voice-to-Task | *Flutter, FastAPI, OpenAI API, PostgreSQL, TickTick API, Core ML*

- Developed a cross-platform voice-to-task assistant (Flutter + FastAPI) with on-device Whisper transcription via Core ML — raw audio never leaves the device — and TickTick OAuth2 sync.
- Applied prompt engineering to convert transcripts into schema-constrained JSON, with an offline queue/retry, async PostgreSQL persistence, Redis rate limiting, and encrypted OAuth token storage.

Mano's Basic Computer — Full Processor Implementation | *Logisim, Digital Logic Design, Assembly*

- Implemented the complete 16-bit processor from Morris Mano's Computer System Architecture in Logisim, replicating all 25 machine instructions with correct opcode decoding and addressing modes.
- Designed the full datapath (ALU, AC, PC, IR, MAR, DR over a common tri-state bus) and a hardwired control unit driving the 3-phase fetch-decode-execute cycle, including indirect addressing and interrupts.

TECHNICAL SKILLS

Languages: Python, C++, JavaScript, TypeScript, Dart, SQL, Assembly (iAPX88)

Frontend: React, Next.js, TypeScript, Vite, TailwindCSS, shadcn/ui, Framer Motion

Backend: FastAPI, Node.js, PostgreSQL, SQLite, Redis, RESTful APIs

AI/ML: LangChain, OpenAI API, Groq API, RAG, Vector Databases (pgvector, ChromaDB), Computer Vision, PyTorch, Ultralytics YOLO, OpenCV, Core ML

Databases: PostgreSQL, SQLite, Firebase, Redis

DevOps & Cloud: Git, CI/CD, Docker, Oracle Cloud, Google Cloud Platform, Render, Vercel

COURSES & ACHIEVEMENTS

CodeRush '26: Runner-up in App Development & Top 5 in ML

Machine Learning Specialization — DeepLearning.AI, Aug 2025

Deep Learning Specialization — DeepLearning.AI, Dec 2025

AI Automation using Make — atomcamp, Apr 2026