

Mohammed Faizan Ahmed

info2faizanahmed@gmail.com | [+91-9391281076](tel:+91-9391281076) | github.com/AhmedFaizanDev | [linkedin.com/in/mohdfaizanahmed](https://www.linkedin.com/in/mohdfaizanahmed) | faizanahmed.in

Summary

AI/ML engineer (B.Tech Computer Science, 2027) who builds and rigorously evaluates AI systems rather than only calling model APIs. Replicated a learned LLM-routing method trained by evolution strategies and measured it against per-model baselines, a random-routing null model, and an oracle-ceiling diagnostic with 95% confidence intervals. Also built an LLM browser agent on the Chrome DevTools Protocol, a U-Net segmentation pipeline on 5-band satellite imagery, and CV extraction pipelines feeding a production recommender.

Education

Mahindra University (Mahindra École Centrale) 2023 – 2027
B.Tech, Computer Science and Engineering Hyderabad, India

Technical Skills

Languages: Python, TypeScript, SQL
ML & Deep Learning: PyTorch, scikit-learn, XGBoost, NumPy, pandas, OpenCV, CNNs, U-Net, semantic segmentation, CMA-ES (derivative-free optimization), TensorBoard
LLM & Agents: OpenAI, Anthropic, Gemini, Groq, Ollama, LLaVA; prompt orchestration, tool calling, multi-turn agent loops, structured output, Chrome DevTools Protocol, WebMCP, RAG, vector search (Qdrant)
Evaluation & Experimentation: held-out benchmarking, baseline & null-model design, oracle-ceiling analysis, confidence intervals, calibration curves, ROC-AUC, Dice/IoU, token & cost accounting
Data & Pipelines: feature engineering, geospatial/multispectral data loading, video frame extraction, batch ingestion, PostgreSQL, Redis
Infrastructure: Docker, AWS EC2, GitHub Actions CI/CD, Django REST Framework, pytest, Streamlit, Linux/shell

Experience

Artificial Intelligence Intern | *AI Linc Technologies Pvt. Ltd.* Jul. 2026 – Aug. 2026
– Built and shipped generative-AI product features for client work, spanning automated video content generation and long-form document generation pipelines, integrated into production backend services with schema-enforced structured outputs, retry with exponential backoff, per-provider timeouts and defined fallback behaviour on malformed model output.
– Built an offline evaluation harness scoring generation quality against a labelled 200-prompt set, making prompt and model changes measurable instead of judged by inspection.
– Used that harness to cut per-request inference cost 35% via model tiering (routing simpler requests to a smaller model) and Redis caching of deterministic sub-results, with no measured quality regression on the eval set.

Full Stack Engineer | *AI Linc Technologies Pvt. Ltd.* Jan. 2026 – Jun. 2026
– Integrated ML models into production-grade backend services, owning the serving path end to end: versioned REST contracts, input validation, request batching, timeout handling, and graceful degradation so model failures did not break the API.
– Built 20+ backend features across a Django codebase, from data model and migrations through API design to the contracts consumed by the frontend team.

Full Stack Developer (Freelance) | *Catalyst Connect, Hyderabad* Feb. 2026 – Mar. 2026
– Built the backend for a production exhibition and stall management platform delivered to a paying client (Node.js, Express, PostgreSQL), with JWT authentication, role-based access and concurrent-booking conflict prevention; deployed on Docker with Redis and CI/CD, and used to run 3 exhibitions covering 200+ stalls.

Backend Developer | *YuVi Collab Technology Pvt. Ltd.* Jun. 2024 – Sep. 2024
– Built 25+ Django REST API endpoints, authentication, and background data workflows over PostgreSQL for client-facing systems.

Head of Tech & Design | *Qubit Club, Mahindra University* Jul. 2025 – Jun. 2026
– Led a 6-person technical team, owning the club website and technical delivery for 8 workshops and events reaching 400+ attendees.

Projects

TinyRouter: Learned LLM Routing via Evolution Strategies | *Python, PyTorch, sep-CMA-ES, NVIDIA H200* 2026
– Replicated the TRINITY coordinator from scratch: a frozen **0.6B** encoder maps each question to a 1024-dim vector and a **≈10K-parameter** head selects both which of three open-source LLMs answers and which role it plays (Thinker / Worker / Verifier), across a multi-turn loop of up to 5 turns terminated early by a Verifier accept.
– Trained the routing head with **separable CMA-ES**, a derivative-free evolution strategy, against a binary correct/incorrect reward, since the routing decision is discrete and non-differentiable and gradient methods do not apply. Added supervised warm-starting from per-(question, model) correctness labels and shaped the training fitness while keeping the evaluation metric strictly binary.
– Evaluated on **120 held-out questions per task** with every single-model baseline averaged over 3 runs to suppress sampling noise, and benchmarked against *random routing* as a null model: **0.858 average vs. 0.835** for the best single model and 0.833 for random; MMLU **0.925 vs. 0.875** random.
– Built an **oracle-ceiling diagnostic** with 95% confidence intervals to test whether the model pool left any headroom to route around, proving math had **+0.049 [0.005, 0.085]** of achievable gain the trained router failed to capture, reframing a flat result as a limitation of the router rather than of the pool and directing where further effort was worth spending.

- Tracked \$62 of API spend exactly from token ledgers; covered the training and grading harness with 54 offline tests.

Autonomous Web Agent | *Python, Chrome DevTools Protocol, Multi-provider LLMs, OAuth 2.0*

2026

- Built an LLM agent that navigates real websites, fills forms and downloads files, representing each page as a **semantic tree built from live CDP DOM traversal** rather than reconstructed from XPath, preserving true parent-child structure, ARIA roles and text wrapped in inline elements that flat element lists lose.
- Implemented **OAuth 2.0 Authorization Code + PKCE** end to end (local callback server, code-to-token exchange, bearer-token injection into browser requests, disk-persisted tokens with silent refresh), so authenticated workflows run without storing passwords or repeating logins.
- Abstracted **8 LLM providers** behind one interface and added WebMCP support for runtime discovery of site-exposed tools; benchmarked a **64% task success rate** over 50 self-defined tasks across 10 websites.

Glacier Segmentation on Multispectral Satellite Imagery | *PyTorch, U-Net, Sentinel-2*

2024

- Trained a U-Net for binary semantic segmentation of glacier boundaries on **5-band Sentinel-2** imagery, writing a custom geospatial loader that tiles multispectral GeoTIFFs into 1,200 training patches and normalizes each band against training-set statistics.
- Reported **Dice 0.87 / IoU 0.78** on spatially disjoint held-out regions rather than pixel accuracy, which is uninformative at this class balance ($\approx 12\%$ glacier pixels); compared BCE, Dice and combined weighted losses and analysed operating points via ROC.

Fault-Tolerant LLM Content Generation Pipeline | *TypeScript, Node.js, Express, OpenAI, Docker, AWS EC2*

2025

- Built a staged generation pipeline producing ≈ 250 -page structured documents from one topic through **≈ 186 sequenced LLM calls**, chaining context via micro-summaries rather than full history to hold coherence inside a bounded context window.
- Added checkpoint/resume at unit granularity so a failure at call 150 resumes without regenerating completed work; held marginal cost to **$\approx \$0.12$ per document** via model tiering. Deployed on AWS EC2 with Docker Compose.

ivm: Incremental View Maintenance Engine | *Python, SQL, DBSP, pytest*

2026

- Built a library implementing the DBSP model that keeps SQL views correct under inserts and deletes by applying only deltas: **37K updates/sec** at p99 43 μ s, **$\approx 3,600\times$** faster than full recompute, verified by **570 differential tests** against an independent recompute oracle; directly applicable to feature stores and incremental data pipelines.

Achievements

Winner: SUDHEE 2025 Hackathon | *CBIT, Hyderabad*

Mar. 2025

- Built a real-time speech-to-Indian-Sign-Language translator, placing first among 60+ teams; owned the streaming speech-to-text pipeline and gesture rendering, tuning it to hold end-to-end latency under 1.2s so output tracked live speech.

3rd Place: AminoVerse Health-Tech Hackathon | *Mahindra University*

Apr. 2025

- Built the Qdrant-backed vector retrieval and backend data-integration layer over four bioinformatics databases (PDB, UniProt, StringDB, DrugBank) for an LLM gene-disease analysis assistant, delivered in a 48-hour build.

Certifications

Oracle Cloud Infrastructure 2025 Certified Generative AI Professional (Oct. 2025) | LLMs, RAG pipelines, semantic search, vector databases.

Red Hat Linux Administration III: Automation with Ansible (Udemy, Feb. 2025) | Linux system administration, shell scripting, Ansible automation.