SAURABH DWADASH SHRENI

Bengaluru, India | +91 79055 38125 | saurabhshreni@gmail.com | **LinkedIn** | **GitHub** | **LeetCode**

SUMMARY

Final-year B.Tech CSE (AI) student who built a multi-agent LLM code review service as an AI Engineering Intern at HCLTech. Trains and evaluates PyTorch models (U-Net 0.85 Dice, EfficientNet-B0 0.879 ROC-AUC) and builds RAG and LLM services in Python, with tests in CI and Docker-based deploys.

EXPERIENCE

AI Engineering Intern **Jun 2026 – Aug 2026**

HCLTech *Certificate* | *Live Demo* | *Code*

- Built a multi-agent AI code review service on ASP.NET Core (.NET 10): 5 LLM agents (bugs, security, explanation, fixes, onboarding) run in parallel via Task.WhenAll, and a supervisor agent merges them into one scored report.
- Put every model call behind one IChatClient layer, so the same 8 REST endpoints run on Groq (Llama 3.3 70B) or local Ollama (Qwen2.5 3B) with no change to calling code.
- Numbered source lines before each LLM call and required every finding to cite `Line [N]:`, turning vague model output into jump-to-line comments a reviewer can check.
- Shipped a PR diff review endpoint (APPROVE / REQUEST CHANGES with a must-fix list), with Docker, GitHub Actions and auto-deploy to Render.

TECHNICAL PROJECTS

GraphRAG++ | *Python, FastAPI, Pydantic, NetworkX, PostgreSQL, Prometheus, pytest* *Code*

- Built a document question-answering service with 11 REST endpoints that fuses vector, BM25 and knowledge-graph search with weighted scoring before the LLM writes an answer.
- Mapped failures to typed states (NO_EVIDENCE, LOW_CONFIDENCE, LLM_FAILURE) so the API returns a warning or partial answer instead of a confident wrong one.
- Added a Prometheus /metrics endpoint with per-stage latency and error counters, and GitHub Actions CI running ruff, mypy and 82 pytest tests on Python 3.11 and 3.12.

MelaNet | *Python, PyTorch, torchvision, OpenCV, Albumentations, Streamlit* *Live Demo* | *Code*

- Built a two-stage PyTorch pipeline for skin lesion analysis: U-Net segmentation on ISIC 2018 (0.85 Dice) and EfficientNet-B0 classification on ISIC 2020 (0.879 ROC-AUC).
- Trained with mixed precision (AMP/fp16) on a single RTX 3050 with Albumentations augmentation, and added Grad-CAM heatmaps so each prediction shows which image region drove it.
- Ran EDA and model evaluation in a Jupyter notebook, with plots for both pipeline stages.

TurboQuant $\times$ **ViT** | *Python, PyTorch, Streamlit* *Live Demo* | *Code*

- Ported LLM KV-cache compression (PolarQuant, TurboQuant) to a Vision Transformer: 10.7$\times$ smaller KV cache and 89% memory saved at 3 bits, holding 83.98% top-1 accuracy.
- Cut cold-start latency from $\sim$40s to $\sim$2s with a process-wide compressor cache and timeout-based fallbacks.

TECHNICAL SKILLS

Languages: Python, C#, Java, SQL, JavaScript

AI and ML: PyTorch, OpenCV, LLM APIs (Groq, Ollama), RAG, multi-agent orchestration, prompt engineering

Data and Backend: FastAPI, ASP.NET Core, REST APIs, PostgreSQL (Supabase), Power BI, Excel, NetworkX

DevOps and Testing: Git, GitHub Actions (CI/CD), Docker, pytest, Prometheus, Render

Certifications: IBM Data Analytics (93%), MathWorks Computer Vision (96.8%), OOP in Java (UPenn)

EDUCATION

Manipal Institute of Technology, Bengaluru (MAHE) **Bengaluru, Karnataka**

B.Tech, Computer Science and Engineering (Artificial Intelligence) | CGPA: 7.69/10 *2023 – 2027*

- Coursework: Data Structures and Algorithms, DBMS, OOP, AI and ML, Agile and Scrum (elective)
- ISC (XII): 83.4% | ICSE (X): 92.4% | No active backlogs | LeetCode: 100+ problems solved