Akshat Balyan

+91 6397743985 | akshat_balyan@outlook.com | linkedin.com/in/akshat-balyan | github.com/B4K2 | b4k2.github.io

EDUCATION

B.Tech in Computer Science and Engineering (AI & ML) 2023 – 2027
Ajay Kumar Garg Engineering College, Ghaziabad

EXPERIENCE

Software Engineer Intern, AMRR Techsols Oct 2025 – Sep 2026

- Built and deployed real-time computer vision systems for factory label-legibility and burner inspection, processing 3 synchronized camera streams at 30–60 FPS with $<$100 ms inference latency.
- Optimized on-site inference pipelines to process continuous video streams without dropping frames.
- Built backend services in FastAPI and Django for a client-facing platform, turning client requirements into modular, low-latency APIs.

PROJECTS

Riftguard – Safety Layer for Parallel AI Coding Agents [GitHub] PyPI package Python, MCP, SQLite

- Built a tool that stops parallel Claude Code / Codex agents from silently contradicting each other: agents share architectural decisions through MCP, and a separate LLM judge flags real conflicts for the developer.
- Records every agent tool call in a content-addressed store and checkpoints automatically before irreversible actions; released on PyPI with CI via GitHub Actions.

Aura – Agentic SRE Co-pilot [GitHub] [Live] LangGraph, FastAPI, React, K8s

- Built an agent that detects incidents, links them to recent deploys, recalls similar past incidents (Qdrant + Neo4j memory), and proposes fixes such as rollbacks and restarts, with human approval required for any destructive action.
- Shipped a terminal UI, web dashboard, and failure-injection demo environment; runs on Docker Compose or Kubernetes.

HackRx 6.0 – Multi-Modal RAG System [GitHub] Top 35 of 10,000+ teams

- Built a RAG API that answers questions over textbook- and legal-scale PDFs on the fly, without pre-computed embeddings; combined AWS Titan and Gemini with token budgeting to keep latency low.
- Containerized with Docker and deployed on AWS App Runner with CI/CD through GitHub Actions.

Mnemos-DNA – Long-Context DNA Language Model [GitHub] Faculty-advised team research project

- Added a k-mer fast-weight memory to linear-attention (KDA) DNA models, restoring exact sequence recall out to 64 kbp for $\sim$3% extra parameters.
- Ran 1 Mbp-context inference in 5.6 GB VRAM on an 8 GB laptop GPU; released weights on Hugging Face and a companion benchmark (recall-bench) on PyPI.

CodeMind – LLM Training on Consumer Hardware [GitHub] Faculty-advised team research project

- Trained 24 GPT-style 60M-parameter models from scratch on 1B tokens to compare attention variants (MHA/MQA/GQA/MLA), RoPE, MoE, and the Muon vs. AdamW optimizers over 250+ GPU hours.
- Built a memory-efficient PyTorch pipeline (gradient checkpointing, chunked cross-entropy, hybrid Muon + AdamW) to train a 1.5B model on an 8 GB RTX 5060 at 1,940 tokens/s.

WhatsApp AI Agent (MCP) [GitHub] 11 stars, 4 forks Google ADK, Go, Python

- Built an autonomous WhatsApp agent with Google ADK and Model Context Protocol: a Go messaging bridge plus a Python MCP server exposing tools for multi-step tasks.

SKILLS

AI/ML: PyTorch, Hugging Face, LLM training & fine-tuning, RAG, AI agents (LangGraph, MCP, Google ADK), MoE, YOLOv8
Backend & Cloud: FastAPI, Django, REST APIs, Docker, Kubernetes, AWS (App Runner, ECR, EC2), GitHub Actions
Data: PostgreSQL, SQLite, Qdrant, Neo4j, Pinecone, ChromaDB, PyArrow
Languages: Python, Go, SQL, JavaScript/TypeScript

ACHIEVEMENTS

Top 35 Finalist, Bajaj Finserv HackRx 6.0 (10,000+ teams) | **3-Star** on CodeChef | **NVIDIA DLI:** Fundamentals of Deep Learning