Abishai K C

Ayapakkam, Tamil Nadu, India · +91 86673 31224 · abishaioff@gmail.com
linkedin.com/in/abishai-k-c-6a5288271 · github.com/Abishai95141 · abishai-ten.vercel.app

Executive Summary

AI engineer with two years of applied research and development experience across LLM systems, agentic tooling, and observability infrastructure. Shipped an open-source MCP server to npm, a deterministic Kubernetes observability engine, and a causal-inference platform.

Skills

- **AI/ML**: LLMs, RAG, Fine-tuning, Model Context Protocol (MCP), Agentic Systems, Evaluation, Causal Inference, Applied ML/DL, Quantization, PEFT/LoRA, ONNX
- **Languages & Frameworks**: Python, SQL, FastAPI, Protobuf, REST, gRPC, PyTorch, TensorFlow, Scikit-learn, LangChain, LangGraph, DoWhy
- **Tools & Infrastructure**: Git/GitHub, Docker, Kubernetes, PostgreSQL, SQLite, FAISS, Ollama, CI/CD, Google Cloud Platform

Professional Experience

- **AI Engineering Intern**, *Matrix Maven – Enterprise AI*
  April 2026 – July 2026
  - **Coodra — shipped to npm as `@coodra/cli` (MIT, public beta)**: Engineered and released a CLI that **removes the re-explanation cost at the start of every AI coding session**: project context is loaded when a session opens, decisions and file changes are recorded while the agent works, and both stay queryable from any later session. Runs with **Claude Code, Cursor, and Windsurf**.
  - **Tooling and policy layer**: Exposed the system as **15 MCP tools** plus a hook bridge that checks an **allow/ask/deny policy before every file write and shell command**, writing an append-only audit log and failing open under load.
  - **Storage and release engineering**: Shipped **offline-first on SQLite** with an opt-in team mode that syncs to a customer-owned PostgreSQL database; held the release gate with unit tests across 9 workspaces and an end-to-end suite that **boots real PostgreSQL in CI on every pull request**.
  - **Model optimization**: Converted and quantized models to the MLX formats for local and Apple Silicon inference.

- **Research Intern**, *GNS-AI, LLC – Supply Chain & Health Tech*
  July 2024 – April 2026
  - **LLM Adjudicator Evaluation**: Measured an LLM adjudicator against a policy reference on **166 prior-authorization cases** across 3 decision classes, reporting **63% agreement**, the direction of its errors, and the finding that **stated confidence did not correlate with accuracy**.
  - **EHR Opportunity Discovery**: Built a prioritization framework covering **17 clinical roles and 4 core hospital workflows** (admission, medication reconciliation, lab order-to-result, referrals), ranking each opportunity on impact, urgency, and feasibility with market validation to **decide which AI product to build first**.
  - **Persona-Based Analytics**: Specified a retail and e-commerce dashboard mapping **10 stakeholder roles** to the KPIs and charts each role acts on, **cutting the reporting surface to the metrics tied to a decision**.

Projects

- **Armature** · Multi-tenant agent platform · github.com/Abishai95141/AI-harness
  July 2026
  - Agent platform where each workspace writes its own tools, verifies them by running them against real data, and keeps the ones that pass, so **repeat work reuses a saved tool instead of regenerating one**. On the same task shape, a cold run took 115.1s and 8,866 tokens; the warm path finished in 9.1s on 964 tokens — **12x faster with 89% fewer tokens** and zero frontier-model calls.
  - Kept the model's own self-test from acting as the gate: **harness-side oracles** check exit status, results, failure sentinels, and output envelope before a tool is stored, with up to 3 execution-feedback repair attempts. Graded against independent oracles — live USGS earthquake counts, a 2,568-order settlement month in which **all 6 injected faults were caught with zero false positives**, and churn scoring on the 7,043-row IBM Telco dataset, with the model trained on a live T4 GPU (**ROC-AUC 0.8345**, 25/25 row checks).

- Executed generated code in **per-workspace Firecracker microVMs with default-deny egress** and per-execution secret brokering; metered every model call against an **up-front USD budget per workspace**; exported stored tools as signed bundles re-executed in the importing tenant before being trusted.
- **Stack**: TypeScript, Node.js, Fastify, React, PostgreSQL (23 migrations), E2B (Firecracker) microVMs, Modal GPU, Model Context Protocol, Docker

- **Vigil** · Kubernetes observability engine · github.com/Abishai95141/Vigil-Dilijens
  April 2026
  - Observability engine for dependency-heavy Kubernetes clusters. Holds measured readings, model forecasts, and human-authored graph notes as separately labelled classes and combines them only at the output layer, so an operator can **decompose any finding back into its evidence** instead of trusting an unexplainable alert.
  - Made findings reproducible by restricting the detection path to **3 arithmetic primitives** (threshold, rate, co-occurrence) with no learned per-cluster state: every evaluation cycle emits a SHA-256 digest, and a replay engine re-runs recorded inputs through the real components and fails loudly on any mismatch. Each of **21 falsification tests** targets one stated guarantee.
  - Added a **TimesFM 2.5** forecasting service behind a string-free wire contract, so the model receives no labels or topology and any conformant model is a drop-in replacement; wrote `graphlint`, a schema and release-immutability validator gating changes to the **842-node ontology**.
  - **Stack**: Go, Python, TypeScript, React, gRPC, Protocol Buffers, Model Context Protocol, Kubernetes, Docker
- **CampaignIQ** · Causal analytics for public health · github.com/Abishai95141/CampaignIQ
  October 2025
  - Measures whether a public-health campaign **changed behaviour**, instead of reporting correlation between exposure and outcome. Estimates effects with an **AIPW doubly robust estimator under K-fold cross-fitting** and reports average and per-segment treatment effects (district, age, existing condition), so spend can be moved toward the groups that responded.
  - Turned segment-level estimates into plain-language allocation guidance with a **locally hosted Mistral-7B**; deployed the React/Flask stack on **IBM LinuxONE (s390x)** with PM2 behind an Nginx reverse proxy.
  - **Stack**: Python, Flask, React, TypeScript, Tailwind CSS, Nginx, Ollama, Causal Inference (AIPW, ATE/CATE), IBM LinuxONE, Docker

Leadership & Activities

- **Assistant Coordinator**, *Tech Society – Saveetha Engineering College*
  August 2025 – Present
  - **HackHustle 2.0**: Led the college's flagship hackathon (May 2026) and grew registrations from **1,600 to 4,500+**. Developed the ERP platform with role-based access control that ran RSVPs, participant queries, and event operations, plus an AI evaluation system using rubric calibration, pairwise comparison, re-ranking, and relative grading that **scored 1,500+ submissions consistently**.
  - **Community Growth**: Ran **5 technical communities** (Machine Learning, Cybersecurity, Game Development, Intelligent Systems, Web Development) and grew active membership from **400 to 600+** by restructuring how sustained participation was rewarded.
- **NASSCOM Conference**: Presented a research thesis on context management in lightweight, on-premises language models, addressing the quality loss that small context windows cause on long documents with a method that summarizes and re-plans mid-document.
- **Breaking the Brain Workshop**: Led a hands-on small language model (SLM) fine-tuning session for 70+ students, taking them from dataset prep to a working fine-tuned model.

Hackathons

- **ABB-Accelerator (top 6 of 1,700+ teams)**: Built a Kubernetes observability tool and presented it to ABB leadership.
- **Gidy Techpreneur (1st of 50+ teams)**: Trained an ensemble model forecasting GDP and recommending budget allocations.
- **Next-Gen, NIT Trichy (1st of 100+ teams)**: Solved a vision-language model problem statement in 24 hours.
- **Fusion for Future (2nd of 60+ teams)**: Shipped a blockchain FIR (police first information report) filing portal with anomaly detection and low gas cost.

Education

- **B.Tech in Artificial Intelligence & Machine Learning; 6 semesters completed**, *Saveetha Engineering College*
  Chennai, India · May 2023 – May 2027