

Shreeshailya Vishwasrao Patil

LinkedIn | ☎ +61 401 325 813 | ♦ Sydney, NSW | ● Portfolio | ✉ shreeshailyp@gmail.com | ● GitHub

PROFESSIONAL SUMMARY

ML engineer shipping production GenAI systems end-to-end. Strong foundation in object-oriented programming (Python, C++, Java) and a track record of performance-first engineering, Redis query-path caching, FAISS index tuning, vLLM quantized inference. Currently pursuing a Master of Artificial Intelligence at UTS (Sydney), drawn to systematic, rigorous engineering teams where algorithmic thinking and software performance intersect.

TECHNICAL SKILLS

AI / ML: Pandas, NumPy, scikit-learn, LangChain, LangGraph, multi-agent frameworks (LangGraph, CrewAI-style orchestration), vLLM, FAISS, PyTorch, TensorFlow, RAG, Generative AI (Claude, OpenAI API), Federated Learning, Differential Privacy, YOLO, Claude Code sub-agents, recommender systems (hybrid content + collaborative filtering)

Languages: Python, C++, Java, JavaScript (ES6+), TypeScript, SQL, Solidity, Shell / Bash

Frameworks: Node.js, Express.js, Next.js 14 (App Router), React, Angular.js, Flask, FastAPI, Django

Databases: Relational Databases (PostgreSQL incl. Supabase, MySQL), Non-Relational (MongoDB, Redis for caching and rate limiting), FAISS (vector search)

DevOps and Cloud: Linux, Docker, Kubernetes, CI/CD (Concourse CI, GitHub Actions), Airflow, Git; AWS (S3, Glue, Lambda, EC2, SageMaker, Step Functions), GCP, Azure, Vercel, Sentry

Observability: Prometheus, Grafana, SLI / SLO definition, structured logging

Systems and Performance: Low-latency API design, high-throughput data pipelines, OOP design patterns, algorithmic problem solving, performance profiling

Architecture: OOP systems design, REST APIs, API integration with third-party systems, microservices, serverless pipelines, data modeling and SQL modelling, data warehousing, data integration, distributed data infrastructure, edge deployment

Analysis and Reporting: Stakeholder communication and reporting to non-technical audiences, Excel (pivot tables, VLOOKUP / XLOOKUP, formula-driven models), Power BI dashboards, Tableau, Jupyter Notebook, data visualization, exploratory data analysis, KPI and revenue analysis, market research and competitor analysis, forecasting and budgeting support

Blockchain: Ethereum smart contracts (Solidity), IPFS off-chain storage, tamper-evident record design

WORK EXPERIENCE

ML Design Engineer | SELEC Controls Pvt. Ltd.

Navi Mumbai, India · Aug 2024 to Nov 2025

- Designed and deployed a high-throughput REST API (Flask + PostgreSQL) serving real-time vector-search queries via FAISS; engineered for low-latency response under concurrent load, containerised with Docker and orchestrated on Kubernetes for zero-downtime production deployments.
- Implemented Redis caching on high-frequency query paths, reducing API response latency by ~40% under concurrent load via query-path caching and reduced compute overhead.
- Built automated data ingestion and processing pipelines using AWS S3 and SageMaker; applied systematic validation at every stage.
- Architected an air-gapped ML inference system using quantized LLMs (GGUF via vLLM) under strict latency SLAs; applied OOP design patterns throughout the LangChain / LangGraph agent architecture, modular, testable, extensible.
- Integrated Prometheus and Grafana for system-wide observability; defined SLI / SLO thresholds and alerting logic, building the monitoring infrastructure required for reliable always-on production systems.
- Engineered a real-time YOLO-based computer-vision system deployed on edge devices; optimised the inference pipeline for speed and accuracy under constrained compute.

AI Intern | Fishburners

Sydney, NSW, Australia · Mar 2026 to Jun 2026

- Architected an agentic workflow: a CrewAI-style multi-agent pipeline orchestrated through LangGraph with 5 specialist agents (chapter architect, research extractor, content writer, QA reviewer, packager) using tool use across a 9-step orchestration with automated QA, transforming heterogeneous source material (decks, video, PDF) into structured modules at production scale. From proof of concept to production deployment.
- Shipped **Hearth Phase 4**, a live ML recommender (Next.js 14 / Supabase) that ranks grants, accelerators, and VC programs against each user profile via hybrid retrieval + collaborative filtering, with an LLM-backed reasoning layer explaining rankings and surfacing deadline conflicts in natural language.
- Scaled agentic delivery beyond one operator: structured state model across Notion + local files so the multi-agent pipeline could orchestrate content, partner intake, and platform builds at pipeline scale.

Software Development Intern | Capgemini

Airoli, India · Jan 2023 to Mar 2023

- Built a production-ready full-stack service-based web application using Angular.js (frontend) and Node.js / Express (backend); implemented 5+ RESTful API endpoints with structured request validation, middleware error handling, and regression tests before each release.
- Designed and optimised a normalised relational database schema (MySQL) applying data modeling best practices; implemented parameterised queries, connection pooling, and ongoing database management to handle concurrent user requests and prevent SQL-injection vulnerabilities.
- Upgraded the authentication system by integrating PingID MFA, lifting login security compliance by 50%; authored test plans covering auth flows, token-expiry edge cases, and session-management failure scenarios.
- Profiled and tuned Apache server configuration via systematic performance analysis, reducing server outages by 40% and page-load times by 20% through optimised caching headers and connection settings.

Analyst (Virtual Experience) | Tata Consultancy Services

Remote · May 2023

- Conducted multi-dimensional data analytics and revenue KPI analysis using root-cause techniques and statistical methods, translating structured data into actionable insights for strategic business decisions.
- Designed and automated Power BI dashboards with data visualisation and geospatial views for multi-region revenue benchmarking, cutting stakeholder data-retrieval time by 35%.
- Documented findings and delivered analytical reports to senior stakeholders in clear, non-technical communication, influencing data-driven business initiatives across multiple regions.

PROJECTS

TradeOps Console, Live Multi-Agent Ops Workbench | [Next.js 16](#) · [TypeScript](#) · [AI SDK v6](#) · [Google Gemini 2.5 Flash](#) · [Vercel](#)

2026

[KV](#) · [tradeops-console.vercel.app](#)

- Built four production AI agents (Invoice Extractor, Inbox Triager, Compliance Pre-Check, Trade Q&A) using the AI SDK with Google Gemini 2.5 Flash; Trade Q&A performs RAG over an in-repo knowledge base with inline citations.
- Implemented per-IP and per-day rate limiting on Vercel KV with fail-open semantics, structured JSON logging to Axiom with PII redaction and trace ID propagation, and a dual-flavour response cache.
- Shipped with 327 tests across 28 files, strict TypeScript, and a four-check CI gate on main.

Cross-Bank Fraud Detection, Federated Learning + Mixture-of-Experts | [Python](#) · [PyTorch](#) · [Federated](#)

Mar 2026 to Present

[Learning](#) · [Differential Privacy](#) · [XGBoost](#) / [LightGBM](#) / [CatBoost](#) · [Streamlit](#) · [github.com/ShreePatil19/Federated-learning-for-cross-bank-fraud-detection](#)

- Initiated and architected a four-person UTS research project on privacy-preserving fraud detection across simulated banks, and first author on the resulting six-page conference manuscript: wrote the core federated-learning engine (federated clients, aggregation server, tabular fraud model) and the experiment entry point that teammates then extended to further datasets.
- Built the triage decision layer that converts per-bank model scores into an operational action, auto-clear, flag for suspicious activity reporting, or defer to a human analyst, via per-expert probability calibration, a per-bank conformal miss-rate guarantee, a cost-optimal flag threshold, and a budgeted deferral rule under analyst capacity limits.
- Benchmarked six federated-learning algorithms (FedAvg, FedProx, FedNova, SCAFFOLD, Per-FedAvg, FedAvgM) across five transaction datasets including IBM anti-money-laundering and synthetic AML, at three non-IID heterogeneity levels via Dirichlet partitioning, with differential privacy, gradient clipping, and Byzantine-client filtering evaluated per algorithm.

Hearth, Funding Radar + ML Recommender (Phase 4) | [Next.js 14](#) · [TypeScript](#) · [Supabase](#) · [Python scrapers](#) · [Claude](#)

2026

[Haiku](#) · [hearth.fishburners.org](#)

- Hybrid recommender: content-based retrieval over opportunity metadata + collaborative signals from user engagement, ranked against user profile.
- Claude Haiku reasoning layer explaining rankings in natural language, surfacing deadline conflicts, and drafting user-facing nudges, running cheaply inside a Next.js API route.
- Observability stack: Sentry for errors, Slack alerts on high-confidence matches, Recharts dashboards. Vercel deploy.

EDUCATION

Master of Artificial Intelligence | [University of Technology Sydney](#)

Sydney, Australia · 2026 to 2027

Relevant coursework: Neural Networks; Fuzzy Logic and Reasoning Under Uncertainty; Advanced Machine Learning; Deep Learning Systems; Natural Language Processing; Distributed Computing for AI.

B.E. Electronics and Computer Engineering | [Fr. C. Rodrigues College of Engineering, University of Mumbai](#)

India · 2020 to 2024

CGPA 8.49 / 10. Honours project in blockchain systems (Ethereum smart contracts and IPFS). Relevant modules: Statistics and Probability, Linear Algebra, Data Structures and Algorithms, Multithreading, Concurrent Programming, Algorithmic Complexity Analysis.

CERTIFICATIONS AND AWARDS

- **Google Cloud Computing Foundations** (Google, 2024), four-course specialisation: Cloud Computing Fundamentals; Infrastructure in Google Cloud; Networking and Security; Data / ML / AI in Google Cloud.
- **Project Cell, ML Member**, Fr. C. Rodrigues College of Engineering (Jan, Mar 2023): Built ML analysis workflows using Linear / Logistic Regression, K-Means, and Random Forests; optimised data preprocessing pipelines achieving a 15% reduction in model-training time.

WORK RIGHTS

Australian Student Visa (subclass 500), eligible for Temporary Graduate 485 (Post-Study Work) after completing the UTS Master of Artificial Intelligence. Open to employer-sponsored 482 / 186 pathways for permanent roles.