ROHAN PRAVEEN CHAVAN

+1-540-824-9961 | rohanchavan0701@gmail.com | linkedin.com/in/rohanpraveenchavan | github.com/RohanChavan0701 | rohanchavan.vercel.app

EDUCATION

Virginia Tech — Blacksburg, VA
M.S. in Computer Engineering, GPA: 3.71/4.0 — *Aug 2024 – May 2026*
- Coursework: Advanced Machine Learning, Trustworthy ML, Computer Vision, Apps of ML, IoT System Design

K.J. Somaiya College of Engineering — Mumbai, India
B.Tech in Information Technology, Distinction, Honors in AI — *Aug 2020 – May 2024*

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, Java, SQL, C++
AI/ML: LLMs, embeddings, RAG pipelines, evaluation frameworks, agentic systems, LLM fine-tuning (SFT, DPO), prompt engineering, LangChain, LangGraph
ML Engineering: retrieval, orchestration, guardrails, model integration, latency/cost optimization, experimentation, Llama 3.3 70B, Hugging Face Transformers
Backend: FastAPI, Node.js, React, RESTful APIs, microservices, CI/CD, Docker
Cloud & Data: AWS (EC2, Lambda, S3, Bedrock), PostgreSQL, ChromaDB, pgvector, Redis

EXPERIENCE

Graduate Research Assistant — Apr 2026 – Present
Virginia Center for Housing Research (VCHR), Virginia Tech — *Blacksburg, VA*
- Engineered production RAG pipeline using Llama 3.3 70B (Together AI primary, Groq fallback), all-MiniLM-L6-v2 embeddings, LangChain, and ChromaDB with two-pass retrieval and 5-tier geographic/income profile routing, scaling the knowledge base to **6,809 indexed document chunks** serving real housing policy users.
- Built evaluation framework with **86 passing tests** covering corpus ingestion, retrieval accuracy, and citation grounding – verified end-to-end quality across 4 Virginia localities and achieved **citation grounding score of 1.0** across all validation targets.
- Integrated HUD, BLS, and Census BPS federal APIs with automated startup warnings and graceful fallback handling; identified and replaced unreliable BLS QCEW data source with Census building permit data, improving pipeline reliability and data quality for production deployment.

Software Engineer – Amazon Nova Trusted AI Challenge — Jan 2025 – Jun 2025
Virginia Tech (Team HokieTokie) — *Blacksburg, VA*
- Fine-tuned LLMs using supervised fine-tuning (SFT) and direct preference optimization (DPO) on **117K synthetic training examples**, reducing adversarial attack success rate by **46%** and outperforming Claude-3.7 Sonnet and CodeLlama-70B on safety benchmarks, securing **1st place (Tournament 2)** among 10 global teams.
- Engineered adversarial evaluation system with taxonomy-guided prompt construction and automated quality scoring across vulnerability detection and content filtering categories, enabling reproducible model experimentation, regression detection, and iteration at scale.
- Optimized inference latency and cost through model fusion and quantization, deploying dual-expert architecture with sequential filtering and reusable components for production LLM security applications.

Software Engineering Intern (AI/ML) — May 2025 – Aug 2025
AutoUnify — *San Francisco, CA*
- Designed and deployed production ML scoring microservice (FastAPI, PostgreSQL, Vertex AI) monitoring accuracy, latency, and cost across multiple live LLM endpoints, providing model integration observability across a distributed multi-service architecture.
- Reduced manual LLM tuning effort by **25%** by building reusable prompt engineering components and automating CI/CD pipeline with pytest suites and JSON Schema validation across agentic workflows.

Process Automation Engineer Intern — Jan 2024 – Jun 2024
Colgate-Palmolive Global Business Services — *Mumbai, India*
- Deployed 3 production FastAPI applications serving **400+ employees** with **70% automation rate** (200+ hours/month saved); reduced response latency from 30s to under 10s via Redis caching and async I/O handling 1K+ daily interactions.

PROJECTS

LunaFlow | *React, TypeScript, Node.js, FastAPI, LangChain, PostgreSQL, Docker, OAuth 2.0* | GitHub — 2025
- Prototyped and shipped agentic AI assistant with tool-calling across Google Calendar, Google Tasks, and ElevenLabs APIs – handling context management, OAuth authentication, and failure recovery end-to-end – deployed with real users on React/TypeScript/Node.js/PostgreSQL stack.

CareRoute | *FastAPI, PostgreSQL, Docker, AWS EC2, A2A Protocol, JSON-RPC 2.0* | GitHub — 2025
- Built and deployed multi-agent healthcare coordination system in **36 hours** (Codefest 2025, **4th Place out of 60 teams**), implementing Agent-to-Agent (A2A) protocol for autonomous task delegation across 6 microservices with domain-specific tool invocation and graceful failure recovery, deployed to AWS EC2.

ACHIEVEMENTS

Amazon Nova Trusted AI: Selected as 1 of 10 global teams; 1st place (Tournament 2), 2nd place (Tournament 1)
Codefest 2025: 4th Place + Honorable Mention among 50+ teams