Harikeshav Rameshkumar

(513) 848-9642 | r.harikeshav@icloud.com | linkedin.com/in/harikeshav-rameshkumar | github.com/Harikeshav-R | harikeshav.me

Education

The Ohio State University — Dean's List & University Honors (all semesters) **Columbus, OH**
B.S. Computer Science; GPA: 3.9/4.0 *Expected May 2028*

- **Relevant Coursework:** Data Structures & Algorithms, Operating Systems, Systems Programming, Computer Networking, Principles of Programming Languages, Database Systems, Software Engineering, Full Stack Web Development, Intro to Deep Learning & AI

Experience

GE Aerospace **Bengaluru, India**
Software Engineering Intern, AI/FinFlow Team *Jun. 2026 – Aug. 2026*

- Cut per-document processing time **74%** (1,957s → 505s) on **FinFlow**, a production contract-extraction platform (pricing, milestones, business rules), by building its pipeline on **AWS Bedrock + Claude** with streaming, rate-limiting, and content-hash-cached page splitting that also handles 448-page (**$101M**) contracts, past Claude's 100-page limit, with **55%** fewer API calls and **37%** less runtime.
- Made financial-lineage data queryable in plain English and production-ready (`mypy --strict` errors 93 → **0** across 31 modules; tests 29 → **78**) by building an **AI assistant** (**FastAPI**, **Amazon Neptune**) and closing a code-execution flaw.
- Delivered accurate adoption analytics for 100K+ employees (**726K+ logs** in **15s**) by building a pipeline (Python, **Polars**, **DuckDB**) and tracing a defect that had inflated the headline number, correcting the figure leadership presented.

Wayfair **Remote**
AI Automation Apprenticeship *Dec. 2025 – Feb. 2026*

- Reduced trend research time by **70%** by building AI agents in **n8n** that scan Amazon and design blogs to identify emerging product trends before mainstream adoption.
- Delivered competitive intelligence across **50+ products** by developing agents tracking price movements, launches, and reviews to surface whitespace opportunities.
- Enabled real-time decisions for the Rugs team by unifying agents into a Market Intelligence Dashboard with automated pipelines.

Palantir Technologies **Remote**
Technical Fellow *Nov. 2025 – Jan. 2026*

- Built end-to-end data workflows processing **10K+ records** by creating ingestion pipelines, transformations, and Ontology objects in **Foundry**.
- Reduced data discovery time by **50%** by designing interactive apps with Object Tables, filters, and visualizations on real-world data.

Siage Solutions **Bengaluru, India**
Software Engineering Intern *Jun. 2025 – Aug. 2025*

- Cut search time **92%** (4.2s → 350ms) across 100+ manuals (15M+ tokens) by building a plain-English Q&A assistant to replace PDF search (**RAG**: LangChain, Qdrant, OpenAI embeddings); now the team's daily tool.
- Cut duplicate tickets **42%** and resolution time from 14.2 to **4.0 hrs** by building a ticket-triage engine (fine-tuned **BERT**, vector index).
- Boosted token throughput **4.8**× for a real-time sentiment service (2,500+ users) by deploying a **vLLM** microservice on **AWS EKS**.

Indian Institute of Technology, Madras **Remote**
Research Intern, Wireless Networks & Spatial AI Group *Jun. 2023 – Oct. 2023*

- Cut manual site-survey effort **86%** by predicting indoor Wi-Fi signal strength anywhere in a building ($R^2 = 0.975$, 100,000+ synthetic points), benchmarking 6 ML models (incl. GPR, SVR, CNN, LSTM) in **Python/TensorFlow**.

Projects

LEAP | *Distributed LLM Inference Engine* – C++20, SIMD, Linux Kernel, PyTorch

- Ran models too large for one GPU (e.g. **Llama 3 70B**) by splitting layers across a ring of consumer machines (**pipeline parallelism**).
- Shrank model memory **4**× with an **INT8** quantizer; also wrote **SIMD** kernels (AVX2/FMA, ARM NEON, OpenMP) and a zero-copy **Linux kernel module** (Netfilter, `mmap`) that bypasses the network stack.

Penumbra-FHE | *FHE Compiler for Encrypted ML Inference* – Python, Rust, ONNX, pytest

- Enabled AI predictions on **encrypted data** without seeing the plaintext, by building an open-source **FHE compiler + runtime**.
- Guaranteed encrypted output ≡ plaintext output, *bit-for-bit*, via a **behavior-equivalence check** in **CI/CD**, by building a parser that converts **ONNX** models into a versioned intermediate format (Python → **Rust** `tfhe-rs`).

Technical Skills

Languages: Python, **C++**, TypeScript, JavaScript, **Rust**, C, C#, **SQL**
AI / ML & LLMs: Multi-Agent Systems (LangChain, LangGraph), **Fine-Tuning** (BERT / Transformers), PyTorch, TensorFlow, LLM Integration, AWS Bedrock, RAG, Prompt Engineering
Systems & Security: Distributed Systems, SIMD (AVX2 / NEON), Linux Kernel Modules, Multithreading & Concurrency, Compilers / IR, Quantization, Cryptography (FHE, AES, X3DH / Double Ratchet), TCP / Sockets
Backend & Web: FastAPI, Flask, **React**, Node.js, REST APIs, WebSockets, Microservices
Cloud & DevOps: AWS (Bedrock, ECS, Lambda, EKS, SageMaker), **Docker**, Kubernetes, **CI/CD** (GitHub Actions), pytest, `mypy`
Databases: PostgreSQL, MySQL, Redis, Vector DBs (Qdrant, ChromaDB, pgvector), Graph DBs (Amazon Neptune), SQLite

Awards

- **2nd Place, Medpace Track, RevolutionUC 2026** (60+ teams) — *Pulse*, multi-agent clinical-trial safety platform
- **3rd Place, Visa Track, TartanHacks 2026** (Carnegie Mellon, 1,200+ hackers) — *Penny*, AI personal-finance assistant
- **3rd Place, The Token Company Track, NextHacks 2026** (2,000+ participants) — *Distill*, LLM context-compression engine
- **Best AI Hack Runner-Up, HackOHI/O 2025** (800+ participants) — *LeadForge*, autonomous agentic AI sales platform