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 2027*

- **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**
Digital Technology Intern, AI/FinFlow Team *Jun. 2026 – Aug. 2026*

- Worked on **FinFlow**, a production platform that extracts business rules, pricing, and milestone data out of legacy contracts and purchase orders; built the extraction pipeline on **AWS Bedrock + Claude**, cutting per-document latency **74%** (1,957s → 505s) through request streaming and rate-limiting.
- Engineered an **AI assistant** that lets finance users query a financial-lineage **graph database** in plain English instead of writing code (**FastAPI** + **Amazon Neptune**); hardened it for production — `mypy --strict` errors 93 → **0** across 31 modules, removed an `exec()` RCE vulnerability, and grew the **pytest** suite from 29 to **78 tests**.
- Extended the pipeline past Claude's 100-page request limit to handle 448-page (**$101M**) contracts via a **page-split-and-merge algorithm** with content-hash caching, cutting API calls **55%** and runtime **37%**.
- Architected an adoption-analytics pipeline (Python, **Polars**, **DuckDB**) processing **726K+ logs** across 100K+ employees in **15s**, and traced a data-integrity defect that had inflated the headline number, correcting the figure leadership presented.

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

- Developed an internal knowledge assistant that lets engineers ask questions in plain English across 100+ technical manuals (15M+ tokens) instead of searching PDFs, using **RAG** (**LangChain**, **Qdrant**, OpenAI embeddings); cut search latency **91%** (4.2s → 350ms) and became the team's daily documentation tool.
- Designed an automated ticket-triage engine that classifies and routes enterprise support tickets, fine-tuning a **BERT** classifier (PyTorch) against a vector index to cut duplicate tickets **42%** and accelerate Mean Time to Resolution to **4.0 hrs** (from 14.2).
- Deployed the real-time sentiment service behind these tools as a **vLLM** inference microservice (**AWS EKS**, Kubernetes, Docker) streaming to 2,500+ daily users, boosting token throughput **4.8**×.
- Scaled a distributed data-ingestion engine (Python **AsyncIO**, Playwright, **Celery**, **Redis**) across a 15-node proxy cluster, collecting **50,000+** reviews/day at **1,200 req/min**.

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

- Modeled indoor Wi-Fi signal strength at any point in a building without physical site surveys, benchmarking 6 ML models (GPR, SVR, CNN, LSTM) in **Python/TensorFlow** to reach $R^2 = 0.975$ accuracy on a synthetic 100,000+-point dataset and cut manual site-survey effort **86%**.

Projects

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

- A **distributed inference engine** that runs models too large for one GPU (e.g. Llama 3 70B) by splitting transformer layers across a ring of consumer machines using **pipeline parallelism**, bypassing the VRAM wall.
- Wrote hand-tuned **SIMD** kernels (**AVX2/FMA**, ARM **NEON**) with OpenMP, a zero-copy **Linux kernel module** (Netfilter, `mmap`) to bypass the network stack, and an **INT8** quantizer shrinking model memory **4**×.

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

- An open-source **compiler + runtime** that runs neural-network inference directly on **encrypted data** via **Fully Homomorphic Encryption**, so a server computes predictions without ever seeing the plaintext.
- Built the front end as a **parser** that lowers **ONNX** compute graphs into a schema-versioned **IR** (Python → **Rust** `tfhe-rs` runtime), gated by a **behavior-equivalence proof** (encrypted output ≡ cleartext, *bit-for-bit*) verified in **CI/CD**.

ArbOS | *Prediction-Market Logical Arbitrage Engine* – Rust, Tokio, watsonx.ai, LangGraph

- A low-latency **Rust** engine that performs **logical arbitrage** on prediction markets: it watches up to **1,000** live **Polymarket** order books and, the instant prices violate the axioms of probability, fires simultaneous **delta-neutral** orders (guaranteed at resolution) at ∼**14 ms** detection.
- Engineered an **LLM reasoning "Brain"** (**IBM watsonx.ai**, **LangGraph**, **NetworkX**) that autonomously maps the hidden logical graph linking markets to drive **4** arbitrage strategies, hardened by **Fill-or-Kill**-only execution, circuit breakers, and exact `rust_decimal` fixed-point money math.

Atlas | *Local-First AI Job-Application Co-Pilot* – Python, Textual, Typer, LiteLLM, SQLite, pytest

- A **local-first**, **terminal-native** job-application co-pilot that discovers matching roles, **AI-scores** each for fit with an explainable rationale, and tailors a **one-page** resume and cover letter per posting from a single master resume — spanning a **Textual** TUI, a **Typer** CLI, and a background **APScheduler** daemon over one shared **SQLite** (WAL) database.
- Architected a pluggable **"bring-your-own-AI"** provider layer behind one `LLMProvider` protocol — unifying coding-CLI backends (**Claude Code**, OpenAI Codex) and **100+** hosted models via **LiteLLM** with a **five-rung** structured-output recovery ladder and availability-only **failover** — all held to **100% line + branch coverage**, `mypy --strict`, and a **9-cell** OS × Python **CI** matrix.

Technical Skills

Languages: Python, **C++**, TypeScript, JavaScript, **Rust**, C, C#, **SQL**
AI / ML & LLMs: Multi-Agent Systems (**LangChain**, **LangGraph**), LLM Integration (**Anthropic Claude**, **OpenAI GPT-4o**, **Google Gemini**), **AWS Bedrock**, **RAG**, Fine-Tuning (BERT / Transformers, PyTorch), TensorFlow, 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