Harikeshav Rameshkumar

rameshkumar.8@osu.edu | harikeshav.me | github.com/Harikeshav-R

Education

The Ohio State University Columbus, OH
B.S. in Computer Science; GPA: 3.9/4.0 *Expected May 2028*
University Honors; Dean's List (all semesters).
Selected coursework: Operating Systems, Systems Programming, Computer Networking, Principles of Programming Languages, Introduction to Deep Learning & AI.

Research Interests

Efficient ML systems and AI infrastructure; distributed inference; LLM and AI-agent runtimes; performance evaluation; privacy-preserving inference.

Independent Research & Systems Projects

Penumbra-FHE — Encrypted ML Inference May 2026 – Present
Independent project *Python, Rust, ONNX*

- Investigating latency–memory–fidelity tradeoffs between radix-integer TFHE and leveled CKKS; built a shared quantized intermediate representation (IR) and integer reference to control graph structure, quantization, and correctness comparisons.
- Evaluated 14 workloads across 13 distinct graphs on an Apple M3 Pro, using fixed worker counts, Criterion latency measurements, and fresh-process peak-memory profiling.
- Observed exact TFHE agreement with the integer reference on 30 checked samples per workload; CKKS ran faster on 11 of 12 supported workloads, with higher peak memory.
- Identified a gap between numerical error bounds and prediction fidelity: CKKS satisfied its predeclared error bound yet changed 80.28% of reference labels on Branched Digits.

LEAP — Heterogeneous Distributed LLM Inference Jan. – Mar. 2026
Independent project *C++20, C, SIMD*

- Explored transformer inference across memory-constrained, heterogeneous machines through a C++20 runtime with asymmetric layer partitioning and activation exchange over a logical Ethernet ring.
- Implemented blockwise INT8 weights, memory-mapped loading, and AVX2/FMA and ARM NEON kernels to address weight storage and cross-architecture execution.
- Built TCP/UDP and Netfilter/shared-memory transport backends, separate prefill/decode protocols, and configurable layer boundaries to explore communication-path design.

Research Experience

Indian Institute of Technology Madras Remote
Research Intern; supervisor: Prof. Venkatesh Ramaiyan *Jun. – Oct. 2023*

- Completed a scoped, supervised study of regression methods for indoor received signal strength (RSSI) prediction using synthetic wireless-propagation data.
- Generated 1,000 samples with a log-distance path-loss model; compared decision trees, Gaussian process regression (GPR), KNN, linear regression, and SVR on a shared 80/20 train–test split.
- GPR achieved $R^2 = 0.975$, MAE $= 0.282$ dB, and RMSE $= 0.931$ dB on the synthetic holdout; documented the need for further validation beyond the simulated environment.

Manuscripts & Technical Reports

1. **Harikeshav Rameshkumar**. *A Controlled Comparison of tfhe-rs Radix-Integer and poulpy-ckks Inference under a Shared Quantized IR.* Unpublished manuscript, work in progress, 2026.
2. **Harikeshav Rameshkumar**. *Machine Learning-Based RSSI Prediction and Heatmap Generation.* Unpublished internship report, IIT Madras, 2023. Supervisor: Prof. Venkatesh Ramaiyan.

Selected Industry Experience

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

- Optimized LLM-based contract extraction on AWS Bedrock and Claude, reducing per-document processing time from 1,957 to 505 seconds through streaming, rate limiting, and cached page splitting.
- Built a FastAPI/Amazon Neptune assistant for financial-lineage queries; expanded tests from 29 to 78 and resolved 93 strict type-checking errors across 31 modules.
- Developed a Python/Polars/DuckDB analytics pipeline processing 726K+ logs in 15 seconds; traced and corrected a join defect that inflated reported adoption figures.

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

- Built retrieval-augmented question answering over 100+ technical manuals (15M+ tokens) with LangChain, Qdrant, and OpenAI embeddings; reduced search time from 4.2 seconds to 350 milliseconds.
- Developed BERT-based ticket triage with a vector index, reducing duplicate tickets by 42%; deployed a vLLM sentiment service on AWS EKS with a reported 4.8× token-throughput increase.

Additional Industry Experience

Wayfair — AI Automation Apprenticeship Dec. 2025 – Feb. 2026
Built n8n agents for product-trend research and competitive intelligence.

Palantir Technologies — Technical Fellow Nov. 2025 – Jan. 2026
Built Foundry ingestion, transformation, Ontology, and data-exploration workflows.

Teaching & Mentoring

Informal Peer Tutor — Computer Science
Tutored peers in Operating Systems, Computer Networking, and introductory Artificial Intelligence and Machine Learning.

Selected Honors & Awards

- **2nd Place, Medpace Track**, RevolutionUC 2026 — Pulse, clinical-trial safety platform.
- **3rd Place, Visa Track**, TartanHacks 2026 — Penny, AI personal-finance assistant.
- **3rd Place, The Token Company Track**, NextHacks 2026 — Distill, LLM context-compression engine.
- **Best AI Hack Runner-Up**, HackOHI/O 2025 — LeadForge, agentic sales platform.

Research & Technical Skills

Languages: Python, C++, Rust, C, SQL.
ML & numerical computing: PyTorch, TensorFlow, scikit-learn, ONNX, quantization, calibration, regression, transformer inference.
Systems: SIMD (AVX2/FMA, ARM NEON), OpenMP, Linux kernel modules, TCP/UDP, memory mapping, distributed execution.
Private inference & evaluation: Quantized IRs, graph lowering, tfhe-rs, CKKS, integer-reference testing, Criterion, latency/memory profiling.
Data & infrastructure: Polars, DuckDB, Qdrant, AWS, Docker, Kubernetes, pytest, GitHub Actions.