Harikeshav Rameshkumar

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

Education

The Ohio State University May 2027
B.S. Computer Science | ***GPA:*** *3.9/4.0* *Columbus, OH*

- **Courses**: Computer Organization, Data Structures & Algorithms, Foundations of Software Engineering
- **Honors**: Dean's List, University Honors

Technical Skills

Languages: Python, C/C++, C#, Rust, TypeScript
Technologies: LangChain/LangGraph, PyTorch, FastAPI, React
DevOps: Docker, Linux, GitHub Actions, AWS, PostgreSQL

Experience

Apprenticeship Jan. 2026 – Present
Pfizer *Remote*

- **Architected an end-to-end document automation platform** for pharmaceutical and mortgage sectors, utilizing **Python, PyMuPDF, and OCR engines (Tesseract, PaddleOCR)** to extract and classify data from massive unstructured "blobs" with custom regex and layout-based heuristics.
- **Developed a high-precision RAG pipeline using LlamaIndex and Gradio**, implementing advanced chunking and metadata filtering to optimize retrieval across open-source models like **Mistral and Phi-2**, resulting in a modular system capable of real-time querying and automated compliance flagging.

Software Engineering Internship Jun. 2025 – Aug. 2025
Siage Solutions *Bangalore, IN*

- **Engineered a scalable, secure RAG architecture and ML ticketing system** that indexed **100+ technical manuals** and automated prioritization for **10+ enterprise clients**; utilized vector databases and semantic embeddings to eliminate ticket redundancy and reduce documentation search time.
- **Developed an end-to-end data intelligence pipeline** to scrape and analyze **50,000+ multi-regional product reviews**, leveraging LLMs for sentiment analysis and thematic extraction to deliver actionable insights via automated trend visualizations and heatmaps.

Research Internship Jun. 2023 – Oct. 2023
Indian Institute of Technology, Madras *Remote*

- **Engineered a predictive ML framework** to model indoor wireless signal propagation using electromagnetic wave theory; conducted large-scale experiments to train models capable of forecasting signal strength across diverse architectural geometries.
- **Authored a comprehensive research paper** evaluating comparative model performance and generating high-resolution spatial heat maps to optimize indoor network planning and deployment strategies.

Projects

LEAP | *C++20, Linux Kernel, LibTorch* 2026

- Engineered a **distributed LLM inference engine in C++20** that pipelines **70B+ parameter models** across consumer devices via a **Ring Topology**, utilizing **INT8 quantization** to solve VRAM bottlenecks.
- Implemented a custom **Linux Kernel Module** for **zero-copy networking** and hand-written **AVX2/NEON SIMD kernels**, achieving **near-native throughput** and enabling real-time inference on heterogeneous hardware.

Distill | *Python, PyTorch, BERT, React* 2026

- Developed an **intelligent LLM context compression library** using a fine-tuned **BERT token classification model** to prune semantic noise, reducing prompt size by **68%** while retaining **99% accuracy** on LongBench V2.
- Built a two-tier filtering pipeline with **word-level aggregation** and **force-token preservation** to prevent hallucinations, achieving **37% lower latency** and **68% cost reduction** for GPT-4o inference.

LeadForge | *Python, LangGraph, React, Docker* 2025

- Architected an **autonomous AI SDR platform** using **LangGraph and Gemini** that orchestrates lead generation and multi-channel outreach, automating the pipeline from Google Maps scouting to **Twilio** voice calls.
- Developed a **generative microservices workflow** that autonomously critiques prospect websites using visual AI and code-generates improved **React/Tailwind prototypes**, deploying them instantly to drive conversion.

Sane Jtreet | *Python, LangGraph, FastAPI, ChromaDB* 2025

- Engineered a **multi-agent trading platform** where specialized LLM agents (Analysts, Debaters, Risk Managers) autonomously collaborate via **LangGraph** to analyze real-time market/sentiment data and formulate investment strategies.
- Implemented a **cognitive architecture with memory reflection** using **ChromaDB**, enabling agents to store past trade outcomes and retrieve "lessons learned" to iteratively improve predictive accuracy and risk management.

Awards

Third Place - Visa Track | *TartanHacks* 2026

- Awarded for **Penny**, a gamified AI-driven personal finance assistant that simplifies expense tracking and financial goal setting through intelligent receipt analysis and interactive mascot-led insights.

Third Place - The Token Company Track | *NextHacks* 2026

- Awarded for **Distill**, a high-performance LLM context compression library enabling drastic reductions in inference costs and latency while preserving accuracy across massive context windows.

Best AI Hack Runner Up | *HackOHI/O* 2025

- Awarded for **LeadForge**, an autonomous AI SDR platform that orchestrates lead generation and multi-channel outreach using LangGraph and Gemini.

Game Development | *World Language Appathon* 2025

- Built an immersive VR market simulator to teach the player Uzbek, featuring a dynamic economy system, physics-based interactions, and AI-driven NPCs, developed for Meta Quest using Unity and the Meta XR SDK.

Various Wins | *Regional Hackathons* 2022 – 2024

- Participated in multiple regional hackathons and won multiple prizes.