# Harish KB

8248052926 | harishkb20205@gmail.com | Harish KB | HARISH20205 | Portfolio

## Education

**Vellore Institute of Technology (VIT)** — Vellore, India
MTech (Integrated) in Computer Science and Engineering(CGPA: 8.59) — Aug 2022 – Jul 2027

## Experience

**AI Engineer Intern** — Jul 2025 – Oct 2025
*Hooman Digital* — Remote

- Pioneered Nosana MCP server for AI assistants to deploy and monitor containerized LLM jobs on decentralized GPU networks with intelligent model-GPU matching and real-time tracking.
- Architected the Inferia ChatHub backend on Nosana Compute, an AI inference platform that dynamically deploys models at scale, implements rate limiting, and streamlines container orchestration, achieving 30% lower inference latency and 40% better resource utilization.

**Generative AI Intern** — Jun 2025
*TITAN Company Limited* — Onsite

- Built an AI-driven fashion visualization pipeline with Runway ML and Streamlined retail ops using n8n, reducing catalog time by 60%, manual effort by 70%, and speeding up product launches by 3x.
- Migrated Taneira's backend from Flask to FastAPI, improving API response time by 40%, scaling throughput by 2.5x, and integrating a modular RAG-based chatbot for AI-driven support.

**AI Research and Development Intern** — Jun 2024 – Oct 2024
*eBramha Techworks Private Limited* — Onsite

- Conducted comprehensive analysis of advanced NLP models like PEGASUS, BERTsum, and BART; applied insights to optimize summarization tasks, improving accuracy by 25% in real-world use cases.
- Developed a speech-to-text system, reducing processing time by 40%, and Constructed an MNIST digit classifier with 95% accuracy using gradient descent and one-hot encoding.

## Research Publications

**FrugalSOT - Frugal Search Over The Models** — 2025
***Co-Author*** *- Accepted at ICICISconf2025 (IEEE)* — Website

- **Conference:** ICICISconf2025 (IEEE Conference Paper), Track 04: Artificial & Computational Intelligence
- Engineered a resource-sensitive model selection process for edge NLP inference which ran on a Raspberry Pi 5 from complexity classification (prompt length, density of NER, syntactic structure), giving a reduction of 21.34% in inference time while sustaining only a 2.68% loss in relevance.
- Devised an adaptive thresholding mechanism driven by exponential moving averages ($\alpha = 0.2$) and cosine similarity–based relevance evaluation, dynamically routing requests across low→mid→high→fallback models using insights from 250 evaluated prompt–response pairs.
- Achieved complete on-device operation through memory-sensitive model selection ($<$8 GB vs $\geq$8 GB RAM), removing cloud reliance while improving latency, data privacy, and real-time performance for IoT and autonomous systems.

## Projects

**Inferia Chathub** | *Python, Docker, SGLang, Nosana, LLM Inference, GPU Optimization* |

- Architected Inferia ChatHub backend on Nosana Compute with integrated SGLang inference runtime for large-scale model deployment, rate limiting, and On-demand scaling, reducing inference latency by 30% across production workloads.
- Optimized inference infrastructure and GPU scheduling mechanisms, minimizing SGLang boot-up time by 45% and increasing tokens-per-second throughput by 40%, ensuring high-concurrency, low-latency inference in production environments.

**SpecQuant** | *Python, Quantization, Speculative Decoding, LLM Inference* | Github

- Innovated SpecuQuant, an adaptive speculative decoding framework leveraging multi-parent quantization (FP16, Q8, Q4), achieving up to 22.6% faster inference on consumer hardware through efficient draft–target token verification and hardware-aware execution.
- Improved response quality via complexity-based prompt classification (syntactic, semantic, and length metrics) ensuring optimal precision selection, maintaining <2% accuracy loss across diverse benchmarks.

**Nosana MCP** | *TypeScript, MCP (Model Context Protocol), Nosana Compute, AI Infrastructure* |

- Designed and implemented the Nosana MCP server, a unified gateway for Nosana Compute, allowing AI assistants to manage decentralized GPU resources, deploy containerized LLM or notebook jobs, and integrate seamlessly with tool-based AI workflows.
- Configured Nosana Compute orchestration through real-time GPU market analytics and adaptive container scaling, delivering 40% faster job execution, 25% higher cost-efficiency, and seamless interoperability with external AI ecosystems.

**PHYDRA** | *Python-C++ Subprocess, FastAPI, ReactJS, Tailwind CSS, Prisma ORM* | Github

- Architected a high-performance Stowage Management System for the International Space Station (ISS) using FastAPI with C++ subprocess orchestration, achieving 95% faster computation and handling 10M+ items in under 5 seconds for real-time stowage optimization.
- Enhanced backend scalability and portability through Docker containerization and Prisma-based database integration, with a React and Tailwind CSS frontend for intuitive monitoring and control.

## Technical Skills

**Languages:** Python, C/C++, JavaScript, TypeScript, Java, Go (Golang)

**AI / ML & DL:** PyTorch, Transformers, LLMs (fine-tuning, inference optimization, quantization), NLP, Computer Vision (YOLO, OpenCV), Scikit-learn

**LLM Systems & Inference:** vLLM, SGLang, Ollama, Eagle, ONNX, Model Context Protocol (MCP)

**Web Development:** React.js, Tailwind CSS, FastAPI, Django, Flask, Express.js

**Databases & Backend:** PostgreSQL, MongoDB, Firebase, Prisma

**DevOps & Cloud:** Docker, NGINX, AWS (EC2, S3), Nosana, Hugging Face Spaces, Vercel, Git, CI/CD

**Systems & Robotics:** Linux, ROS2, Bash, CMake, Raspberry Pi

## Certifications

- **Coursera:** Supervised Machine Learning: Regression and Classification
- **Coursera:** Advanced Learning Algorithms
- **Coursera:** Unsupervised Learning, Recommenders, Reinforcement Learning
- **Coursera:** Generative AI with Large Language Models