Vamshi Nagireddy

(916) 707-2957 — vamshi.knagireddy@gmail.com — San Jose, CA
linkedin.com/in/vamshinr — github.com/vamshinr — medium.com/@vamshire

SUMMARY

- Software Engineer specializing in benchmarking and performance optimization.
- Built telemetry-driven tuning frameworks at Intel that improved application performances by 20-30%
- Benchmarked Stable Diffusion and LLMs across runtimes(GPU, NPU), implemented quantization and RAG pipelines for OpenVINO.
- Deep hands-on with PyTorch, CUDA, and profiling (roofline, Nsight-style analysis).

SKILLS

- **Programming Languages** - Python, C, C++, Java, Shell Scripting, Go, Bash.
- **Machine Learning And AI** - TensorFlow, PyTorch, CUDA, Triton, vLLM, TensorRT-LLM, MLIR, ONNX.
- **DevOps And Cloud Technologies** - Docker, Kubernetes, Jenkins, Git, Kafka, MLFlow.
- **Data Analytics And Visualization** - Tableau, SQL, NoSQL (MongoDB, Neo4j, DynamoDB, Cassandra).
- **Build And Automation Tools** - CMake, Shell Scripting, Clang/LLVM.

EXPERIENCE

Software Engineer — Intel Corporation — San Jose, CA **July 2023 - Present**

- Contributed to the first release of the **MLPerf Client**(Llama2 7B, INT4 weights, FP16 activations), benchmarked across **TTFT** and **token generation rate** metrics across **4 OpenOrca prompt** configurations spanning various tasks.
- Drove **Intel's IHV-format** submission via **OpenVINO** (**GPU**), while coordinating cross-vendor coverage across **ORT-GenAI** with **DirectML, TensorRT-LLM, QNN, and CoreML execution providers**.
- Validated accuracy using **MMLU** on **quantized ONNX reference models** while enabling **IHV-format quantization** path for **Intel maintaining ONNX-RT baseline** for other vendors.
- Architected dynamic runtime optimizer **(user-mode + kernel-mode + IOCTL)** that **tunes thread placement and memory policy** per workload, improving throughput up to **25%** on compute-intensive AI kernels.
- Architected **telemetry-driven runtime optimization frameworks** that dynamically tune execution policies based on workload characteristics, improving application performance by up to 25%.
- Built an LLVM-based **profile-guided optimization framework (HWPGO/PGO)** that uses hardware execution data to generate and apply optimized code variants via transformations such as unrolling, inlining, and vectorization for performance-critical paths.

Machine Learning Intern — Intel Corporation — San Jose, CA **May 2022 - August 2022**

- Developed and evaluated a **benchmarking suite** for assessing **Stable Diffusion** and **LLM text generation** performance across multiple **runtime backends**.
- Implemented a **HWINFO parser** to surface **hardware bottlenecks**, reducing manual analysis time by 50% and **accelerating optimization cycles**.
- Experimented with various **RAG pipelines, prompt engineering, model quantization, and fine-tuning** to optimize performance and adaptability contributing to **Optimization** of **Intel's AI software stack (OpenVINO) for LLM inferencing.**

AI Research Assistant — CSU Sacramento — Sacramento, CA **January 2022 - May 2022**

- Developing state-of-the-art **video-to-story AI system** achieving **71%** accuracy on topic classification (vs **57.5%** baseline.)
- Engineered multimodal pipeline integrating **YOLOv8,WHisper, BLIP-2, and Llama** for video understanding tasks.
- Performed downstream tasks from a curated dataset of 1002 annotated video advertisements for persuasion strategy identification using **LLM** for **zero-shot inference** in marketing videos.
- Achieved **35%** improvement over baselines on 8/9 **long-form video understanding benchmarks**.

Machine Learning Engineer — Phenom — Ambler, PA **July 2019 - July 2021**

- Spearheaded the design and optimization of sequence models (**LSTMs, and Transformers**) for a resume parsing tool, improving Named Entity Recognition accuracy by **10%** and text classification F1-score by **8%**, leading to more precise **data extraction**.
- Developed a job ranking pipeline integrating **ML model** outputs with Elasticsearch indices, tuned relevance **scoring and ranking algorithms**, and surfaced real-time candidate-job match insights via **Kibana** dashboards increasing **user engagement** by **10%**.
- Designed and implemented a **BERT**-based sequence classifier for resume section distinction and raw text extraction from unstructured data, achieving **95%** accuracy in identifying sections like "Work Experience" and "Education."
- Engineered a robust data pipeline integrating **Kafka** with the **ELK stack** (Elasticsearch, Logstash, Kibana) and instrumented microservices with **Prometheus metrics and Grafana dashboards** for **real-time operational visibility**, reducing data processing time by **20%**.
- Managed comprehensive **model training and lifecycle management** using **MLFlow**, reducing **model deployment time by 25%** and ensuring continuous performance monitoring and updates of deployed models.

PROJECTS

- **VaultASR.** Built a high-performance, fully local, and privacy-first speech-to-text pipeline. Intergated with OpenAI's Whisper, this solution integrates advanced Voice Activity Detection (VAD) and Speaker Diarization. It runs entirely on-device with optimized GPU acceleration across multiple execution frameworks, including CoreML, DirectML, CUDA, and ROCm.(Github)

- **PEARL - Proactive Execution and Adaptive Reasoning Loop.** Built an FastAPI based autonomous AI agent with the Gemini API, featuring a PEARL cognitive loop for dynamic task decomposition and execution. Integrated external tools (web search) and a ChromaDB vector database for persistent, long-term memory, enabling experience-based learning and planning.(Github)
- **DrugGuard - An LLM-Powered Drug Side Effect Retrieval System.** Developed a conversational AI system to accurately retrieve drug side effects by implementing and evaluating RAG and GraphRAG architectures. Engineered a data pipeline to process and store 19,520 drug side effect associations, covering 976 drugs and 3,851 side effect terms.(Github)
- **Freight Broker AI Agent - Secure Freight Management API.** Built API with secure endpoints for freight loads and call data, using API key authentication and end-to-end n8n workflow automation. Integrated SQLite for call metrics (sentiment, outcome, negotiations). Designed interactive dashboard and n8n-driven real-time analytics, featuring charts and metrics. (Github)

EDUCATION

Master of Science (MS) in Computer Science — California State University Sacramento 2021 – 2023
GPA: 3.7. Relevant Coursework: DS Algorithms, Machine Learning, Distributed Systems.

Bachelor of Science (BS) in Computer Science — JNTUH College of Engineering Hyderabad 2016 – 2020
GPA: 3.9. Relevant Coursework: Database Systems, Data Structures, AI.

PUBLICATIONS AND CONTRIBUTIONS

- ***"Job recommendation system with NoSQL databases: Neo4j, MongoDB, DynamoDB, Cassandra and their critical comparison."*** Evaluated NoSQL databases for job recommendation, analyzing query times, replication, and consistency. Designed a scalable engine with graph-based algorithms, recommending optimal database choices for user-skill-job mapping. (CSU Sacramento Digital Repository)
- Active contributor to the ***"BAPCo benchmarking consortium"***.
- ***"The Hidden Journey of a Prompt: Unpacking the Bottlenecks of LLM Inference."*** Explores the latency sources across the LLM inference stack, from data loading and attention kernels to memory bottlenecks. (medium)
- ***"From torch.device("cuda") to GPU Hardware: The Hidden World Behind a Single Line of PyTorch Code."*** Comprehensive guide connecting PyTorch device abstraction to CUDA context creation, PCIe DMA and GPU hardware architecture. (medium)