VAISHNAVI PANCHAVATI

Mountain View, CA

vpanchavati10@gmail.com vaishnavi-panchavati vaishdho1 portfolio

Work Authorization: Eligible to work in the U.S with valid **H4 EAD** (no sponsorship required)

Education

University of Massachusetts Amherst — **Feb 2023 – Dec 2024**

*Master of Science in Computer Science **(Minor: Machine Learning)**, GPA : **4.0/4.0*** — *Amherst, MA*

National Institute of Technology Karnataka (NITK) — **Aug 2014 – May 2018**

*Bachelor of Technology in Electrical and Electronics Engineering, CGPA : **9.26/10*** — *Surathkal, Karnataka*

Relevant Coursework: **Advanced Natural Language Processing**, **Neural Networks**, **Advanced Algorithms**, Reinforcement Learning, Distributed Systems, Data Structures and Algorithms

Technical Skills

Languages: C, ***C++***, ***Python***, Go, SystemVerilog, JavaScript, HTML, CSS, Shell Scripting
Frameworks & Libraries: Flask, Django, ***PyTorch***, ***NumPy***, ***gRPC*****, RESTful APIs,** ***TensorFlow Serving*****, vLLM**
Tools & Platforms: ***Docker*****,** ***Kubernetes*****,** ***AWS (EC2)*****, Git, Linux**
Databases: ***PostgreSQL, SQLite, MongoDB***
ML and Systems: Model Distillation, ***Distributed ML Inference*****, Model Serving,** ***Inference Optimization***

Work Experience

Texas Instruments — **Aug 2019 – May 2022**

Software Engineer (5G Transceiver) — *Bangalore, Karnataka*

- Designed and implemented ***C/C++ based functional test frameworks*** for evaluating multiple embedded subsystems, ensuring ***high system reliability*** and precise feature validation.
- Developed ***scalable Python based orchestration*** to manage regression test workflows, ***reducing manual validation time by 40%.***
- Migrated signal-processing logic from MATLAB to modular Python packages and ***integrated it into a versioned local cloud test system, reducing datapath errors by 60%.***

Qualcomm — **Sep 2018 – April 2019**

Software Engineer (4G/5G Small Cells) — *Hyderabad, Telangana*

- Implemented features for LTE/5GNR PHY layer protocols in CATM1 and NBIoT devices in C++ and Python.
- Identified corner cases to ***boost performance by up to 90%*** in key tests like throughput and SINR for both FDD and TDD systems.

Projects

Model Serving System | ***Python, Kubernetes, Go***

- Built a ***distributed LLM*** serving system with ***gRPC and Python***, supporting autoscaling, model caching, fault tolerance, dynamic model loading, and MongoDB backed persistence.
- Deployed the system using a ***custom Kubernetes Operator*** and ***Prometheus*** for real-time performance monitoring.

Stock Bazaar Application | ***Flask, Docker, AWS, Python***

- Built a ***fault-tolerant, scalable microservices*** based stock trading system using ***thread-per-request concurrency and locking*** to handle coordinated client requests.
- Implemented ***server push cache consistency*** and leader based Order replication with container deployment on ***AWS EC2, reducing*** latency by ***5ms***.

Quote-MI | ***Pytorch, LLM***

- Fine-tuned ***BERT, RoBERTa, GPT-2, T5*** on manually curated quote datasets for multi-label classification obtaining ***70% accuracy*** and evaluated T5 and Gemma using zero and few-shot prompting.
- ***Deployed GPU-accelerated BERT*** with TensorFlow Serving + Docker, ***reducing*** inference latency by ***120ms.***

Quantization for model inference | ***Python, GCP***

- Implemented ***GPTQ*** based 4-bit ***quantization pipeline*** with on the fly dequantization, reducing model size with minimal degradation in performance.
- Added custom quantized linear layers and packed state compression, cutting memory footprint by ***50%*** for a 70GB model.

miniLlama | ***Pytorch***

- Implemented core features of a compact Llama 2 model and performed sentence completion with temperature sampling.
- Fine-tuned the model for sentence classification on SST and CFIMDB using a custom LoRA approach, improving CFIMDB performance by ***30%***.