Sudhir Pol

* Seattle, Washington +1 (930) 904-4785 sudhirpol522@gmail.com linkedin.com/in/sudhir-pol-a6b544179 sudhirpol522.github.io

Summary

Machine Learning Engineer working on LLM inference and model optimization, with 3 years of production ML at Adobe, S&P Global, and American Express. Implemented the FlashAttention V2 forward pass in CUDA C++ and made it 8.5 times faster, and quantized and served 7B models on vLLM at 2.2 times faster decode for 56 percent less cost.

Skills

Languages: CUDA C++, C++, Python, SQL, Java

Inference and Optimization: Quantization (AWQ, GPTQ, 4 bit), vLLM Serving, Speculative Decoding, Tiled Attention and Online Softmax, Kernel Profiling, Memory Bandwidth and HBM Traffic Analysis, Warp Level Reductions, Latency and Cost per Token Benchmarking

GPU and Tooling: Nsight Compute, PyTorch and SDPA, llm-compressor, HuggingFace Transformers, PEFT and LoRA, LiteLLM, A100 GPUs

Systems and MLOps: Docker, Kubernetes, ArgoCD, GitHub Actions, CI/CD, AWS (SageMaker, EC2, Lambda), MLflow, DVC, Weights and Biases, FastAPI

Projects

FlashAttention V2 from Scratch | *CUDA C++, Nsight Compute, PyTorch* **Apr. 2026**

- Implemented the FlashAttention V1 and V2 forward pass, tiled attention with online softmax, in CUDA C++ from the papers, matching PyTorch SDPA output to 2e-6 in fp32
- Optimized the kernel 8.5 times by restructuring the tiling loop to remove redundant HBM traffic, eliminating shared memory bank conflicts, and parallelizing the softmax with warp shuffle reductions, profiling each change in Nsight Compute

QuantRoute: Quantization Aware Inference Router | *vLLM, llm-compressor, AWQ, GPTQ, A100* **Nov. 2025**

- Quantized Qwen2.5-7B to 4 bit with AWQ and GPTQ using llm-compressor and served multi precision lanes on vLLM, delivering 2.2 times faster decode at 56 percent lower serving cost on financial tasks
- Trained a difficulty classifier that routed each request to the cheapest lane that could answer it, with an LLM as Judge escalation loop and an offline oracle analysis to locate the routing bottleneck

MultiDocChat: Production RAG System | *FastAPI, FAISS, Kubernetes, ArgoCD* **May 2025**

- Served a retrieval augmented generation system behind FastAPI with FAISS indexing, delivered through GitOps with GitHub Actions, ArgoCD, and Kubernetes for zero downtime rolling deployments

Experience

Adobe **May 2025 - Aug. 2025**

Machine Learning Engineer Intern *San Jose, California*

- Built an LLM as Judge benchmarking harness with LiteLLM that ran Gemini 2.5 Pro, GPT 4.1, and Claude 4.5 behind one interface on multimodal chart quality, raising offline F1 from 79 percent to 92 percent by rebuilding the scoring rubric and tracking runs in Weights and Biases
- Shipped a LangGraph agent that called Adobe PDF Spaces tools through Model Context Protocol with LangSmith tracing and a human approval gate, replacing a 48 hour manual review cycle with a 30 minute pipeline

S&P Global **Nov. 2022 - Aug. 2024**

Data Scientist II *Hyderabad, India*

- Deployed Llama 2 on AWS SageMaker and Lambda with chain of thought prompt templates for extraction across long documents, cutting close to 20 minutes of manual work per analyst task
- Fine tuned and served LayoutLMv3 for token classification on AWS EC2, versioning datasets and models with DVC and MLflow and promoting releases through a CI/CD pipeline so runs stayed reproducible

American Express **Dec. 2021 - Nov. 2022**

Data Scientist I *Gurgaon, India*

- Trained and shipped VIBE, a call quality scoring model combining XGBoost with BERT embeddings, containerized with Docker and run as the production scoring service the annual bonus pool was decided on
- Served a complaint risk classifier and an LSTM intent model behind Flask on AWS EC2, using PySpark for feature generation

Writing

- Speculative Decoding: Make LLM Inference Faster, draft and verify decoding for 2 to 3 times speedup *Oct. 2025*
- Demystifying GPTQ: From Lagrange Multipliers to Vectorized PyTorch *Jun. 2026*
- The Math of AWQ: Protecting Salient Channels from the Inside Out *Jun. 2026*

Education

Indiana University **Aug. 2024 - May 2026**

Master of Science in Data Science *Bloomington, Indiana*

Rajarambapu Institute of Technology **Aug. 2017 - May 2021**

Bachelor of Technology in Electronics and Tele-Communications *Sangli, India*