Sudhir Pol

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

Summary

AI Engineer with 3 years of experience designing, building, and deploying AI models and systems in production for cross functional teams. Covers the full path from data preprocessing and feature engineering to model evaluation, fine tuning for performance and scalability, and continuous delivery on AWS with Docker, Kubernetes, Git, and automated CI/CD pipelines. Master of Science in Data Science, with a habit of documenting systems clearly, monitoring models after release, and troubleshooting deployment issues alongside data scientists and software engineers.

Education

Indiana University **Aug. 2024 – May 2026**
Master of Science in Data Science *Bloomington, United States*

Rajarambapu Institute of Technology **Aug. 2017 – May 2021**
Bachelor of Technology in Electronics & Tele-Communications *Sangli, India*

Experience

Adobe **May 2025 – Aug. 2025**
Machine Learning Engineer Intern *San Jose, California*

- Built an automated evaluation harness with LiteLLM to benchmark Gemini 2.5 Pro, GPT 4.1, and Claude 4.5 on multimodal chart quality tasks, tracked every run, metric, and artifact in Weights and Biases, and raised the offline F1 score from 79% to 92% by redesigning the scoring rubric
- Developed a LangGraph agent on Model Context Protocol (MCP) over Adobe PDF Spaces that cut evaluation feedback from 48 hours to 30 minutes, instrumented with LangSmith tracing for monitoring and a human review gate that let stakeholders approve or reject each generation plan

S&P Global **Nov. 2022 – Aug. 2024**
Data Scientist II *Hyderabad, India*

- Designed, deployed, and maintained a question generation and answering API on AWS EC2 using a T5 transformer model, adding logic that pulled current year figures out of long text documents for downstream systems
- Fine tuned LayoutLMV3 for token classification on financial filings, ran error analysis on a held out sample to measure accuracy by field type, and released the model into a QA environment on AWS EC2 for validation before production rollout
- Deployed Llama 2 with SageMaker and AWS Lambda and applied chain of thought prompting to extract structured fields from paragraphs, saving analysts close to 20 minutes of manual collection per document
- Owned the MLOps workflow for the team with DVC for dataset and model versioning, MLflow for experiment tracking, and GitHub Actions continuous integration and delivery with Docker, which gave reproducible releases and clear documentation for every deployed model
- Engineered a Java regex framework with named groups to parse varied structures from financial documents, which replaced a third party vendor with an in house solution and saved close to 10 minutes of analyst time per filing

American Express **Dec. 2021 – Nov. 2022**
Data Scientist I *Gurgaon, India*

- Designed the VIBE sentiment scoring model that combined XGBoost with BERT embeddings to grade call quality, containerized it with Docker for repeatable deployment, and used the resulting scores to set yearly bonus decisions for customer representatives
- Built a three class risk classification model for customer complaints using a naive Bayes classifier, TFIDF features, and isotonic regression for probability calibration, then deployed and monitored it on AWS EC2
- Developed an LSTM demand prediction model that mapped call transcripts to demand types and served it with Flask on AWS EC2, which let the business route and prioritize incoming customer calls by need
- Processed high volume call transcripts with PySpark, TFIDF, and regex to detect missing offer disclosures, building the preprocessing and feature pipeline that flagged representatives giving incomplete information

Projects

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

- Built and shipped a document question answering service with FastAPI, LangChain LCEL, FAISS indexing, and MMR retrieval, using Gemini 2.5 Pro for semantic scoring and LangSmith for end to end tracing of every request in production
- Implemented GitOps continuous delivery with GitHub Actions, ArgoCD Image Updater, and Kubernetes manifests to reach zero downtime rolling deployments with semantic versioning and automated rollback

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 multiple precision lanes on vLLM, reaching 2.2 times faster decoding at 56% lower serving cost across financial tasks
- Built a learned difficulty classifier and an escalation loop that routes hard requests to the higher precision lane, then used an offline oracle analysis to isolate routing bottlenecks and tune the difficulty signal

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

- Implemented the FlashAttention V1 and V2 forward pass with tiled attention and online softmax in CUDA, then made it 8.5 times faster by restructuring the tiling loop to remove redundant memory traffic, clearing shared memory bank conflicts, and parallelizing the softmax with warp shuffle reductions
- Profiled every revision in Nsight Compute and validated the kernel against PyTorch SDPA output to within 2e-6 in fp32, documenting each optimization step and its measured effect

Blog Posts

- The Math of AWQ: Protecting Salient Channels from the Inside Out | Activation aware weight quantization *Jun. 2026*
- Demystifying GPTQ: From Lagrange Multipliers to Vectorized PyTorch | Optimal brain quantization *Jun. 2026*
- Speculative Decoding: The Math Behind Faster LLM Inference | Parallel draft and verify decoding *Oct. 2025*

Skills

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

AI and Machine Learning: LLMs, Fine Tuning, LoRA and PEFT, Quantization (AWQ, GPTQ), RAG Systems, Agentic Workflows, Model Context Protocol (MCP), Data Preprocessing, Feature Engineering, Model Evaluation, Deep Learning, NLP, Computer Vision

Frameworks: PyTorch, TensorFlow, HuggingFace Transformers, vLLM, LangChain, LangGraph, LiteLLM, Scikit-learn, Pandas, NumPy

Cloud and Deployment: AWS (EC2, S3, SageMaker, Lambda), Docker, Kubernetes, ArgoCD, FastAPI, Flask, REST APIs

MLOps and Practices: GitHub Actions CI/CD, Git, MLflow, DVC, Weights and Biases, LangSmith, Nsight Compute, Model Monitoring, Technical Documentation, Agile and Scrum