Sudhir Pol

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

Summary

Machine Learning Engineer with 3 years of experience taking models from prototype to production at Adobe, S&P Global, and American Express. Owns the full path from fine tuning and evaluation through containerization, deployment, and monitoring on AWS and Kubernetes, and treats latency, cost, and reliability as design constraints rather than afterthoughts.

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*

Experience

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

- Built an LLM as Judge evaluation harness with LiteLLM that benchmarked Gemini 2.5 Pro, GPT 4.1, and Claude 4.5 on multimodal chart quality, raising offline F1 from 79 percent to 92 percent by rebuilding the scoring rubric against a human labeled reference set and tracking every run in Weights and Biases
- Shipped a LangGraph agent that called Adobe PDF Spaces tools through Model Context Protocol, replacing a 48 hour manual review cycle with a 30 minute automated pipeline
- Instrumented the agent end to end with LangSmith tracing and a human approval gate on every generated plan, giving the team a debuggable execution record and a release control that stopped bad plans before they ran

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

- Fine tuned and served LayoutLMv3 for token classification over financial documents on AWS EC2, versioning datasets and model artifacts with DVC and MLflow and promoting releases through a CI/CD pipeline so training runs were reproducible and rollbacks were safe
- 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 data collection per analyst task
- Built a question generation and answering API on AWS EC2 around a T5 transformer, and engineered a Java parsing framework with named capture groups that replaced a paid third party vendor with an in house service, saving roughly 10 minutes per analyst task

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 representative bonus pool was decided on
- Served a complaint risk classifier with isotonic calibrated probabilities and an LSTM intent model behind Flask on AWS EC2, generating features from the transcript corpus with PySpark so incoming calls were routed by predicted need and risk tier

Projects

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

- Built a production retrieval augmented generation service with FastAPI, LangChain LCEL, FAISS vector indexing, and MMR retrieval, scoring answer relevance with Gemini 2.5 Pro and tracing every request in LangSmith to isolate retrieval failures
- Implemented GitOps delivery with GitHub Actions, ArgoCD Image Updater, and Kubernetes manifests, achieving zero downtime rolling deployments with semantic versioning

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

- Served multi precision lanes of Qwen2.5-7B on vLLM after 4 bit quantization with llm-compressor, 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 that located the routing bottleneck

Writing

- Speculative Decoding: Make LLM Inference Faster *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*

Skills

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

Machine Learning: Deep Learning, NLP, Transformers, Large Language Models, Fine Tuning (LoRA, PEFT), RAG and Vector Search, Prompt Engineering, Model Evaluation and LLM as Judge, Probability Calibration, XGBoost, Random Forest

Frameworks and Serving: PyTorch, TensorFlow, HuggingFace Transformers, Scikit-learn, LangChain, LangGraph, LiteLLM, vLLM, Model Context Protocol, FastAPI, Flask

Infrastructure and MLOps: AWS (SageMaker, EC2, Lambda, S3), Docker, Kubernetes, ArgoCD, GitHub Actions, CI/CD, MLflow, DVC, Weights and Biases, LangSmith, Git