SAUMIT PAUL

+1 (984) 335-8991 ◇ Raleigh, NC ◇ sspaul3@ncsu.edu ◇ LinkedIn ◇ GitHub ◇ Website

SUMMARY

ML Engineer with 3+ years of experience shipping production computer vision systems for medical imaging. Currently focused on designing autonomous LLM reasoning workflows and implementing production observability systems.

EDUCATION

North Carolina State University 08/2025 - 05/2027
Master of Science, Electrical Engineering GPA: 4.0
Relevant Coursework: Generative AI For Systems, Data Science for Signal Processing, Neural Networks

Manipal Institute of Technology 07/2018 - 08/2022
Bachelor of Technology, Electrical and Electronics Engineering GPA: 3.8

SKILLS

Programming	Python, R
AI Engineering	LangChain, Langfuse, Ollama, Groq
Machine Learning	PyTorch, TensorFlow, Transformers, Scikit-Learn
Data	Pandas, NumPy, Matplotlib, CVAT, SQL, MongoDB, OpenCV
MLOps	MLflow, Git, DVC, ONNX

PROFESSIONAL EXPERIENCE

Data Scientist 06/2022 - 07/2025
DeepTek.ai

- Developed PyTorch computer vision pipelines for chest X-ray analysis achieving over 90% AUROC for pleural effusion detection in deployed FDA-approved clinical workflows.
- Architected interactive experimentation platform with automated inference, MLflow based tracking and metric computation, reducing model evaluation turnaround from several hours to under 5 minutes.
- Led regulatory validation across US FDA, Thai FDA, HSA, and CE approvals by coordinating MRMC study with 24 radiologists to demonstrate AI system efficacy.
- Maintained MongoDB annotation database for over 1.8M chest X-ray studies and led CVAT migration, ensuring data consistency and infrastructure reliability across clinical datasets.

Automation Intern 01/2022 - 05/2022
Anheuser-Busch InBev

- Designed Power BI dashboard on employee work patterns and built SQL database infrastructure.
- Performed data wrangling on Task Mining API data with over 1M weekly records using Azure Data Pipeline for analytics workflows.

PROJECTS

Floki: MLflow Experiment Agentic Assistant

- Built CLI based tool using LangChain agents to autonomously query and synthesize MLflow experiment data to answer comparative analysis questions without human intervention.
- Integrated Langfuse observability to trace agent execution across tool calls, enabling debugging and monitoring of multi-step reasoning workflows in production.
- Implemented structured tool validation with Pydantic schemas to enforce correct argument passing and prevent common LLM errors in tool invocation.

LLM Based Optimization System for Cache Replacement Policy Search

- Built autonomous agentic LLM-driven optimization workflow using OpenAI API and Ollama to iteratively generate, evaluate, benchmark and refine C++ cache replacement policies across SPEC CPU 2006 workloads.
- Designed RAG-enabled memory system using SQLite and graph-structured policy history to guide autonomous branching exploration and prune weak candidates, reducing evaluation time by 33%.

PUBLICATIONS

- A Comprehensive Evaluation of DeepTek CXR Analyzer in Detecting and Localising Suspicious Findings in Chest X-rays (https://dx.doi.org/10.26044/ecr2024/C-23322)
- Artificial Intelligence as a Proficient Tool in Detecting Pulmonary Tuberculosis in Massive Population Screening Programmes: A Case Study in Chennai, India (https://doi.org/10.2185/jrm.2024-015)