Sudipta **Pathak**

Lead AI Infrastructure Software Engineer ·
Jackson Township, 08527, United States of America
(+1) 848-373-5136 | sudipto.pathak@gmail.com | https://github.com/sudiptap | https://www.linkedin.com/in/sudipta-pathak-a6910117/

Summary

Lead Machine Learning Infrastructure Engineer with 8+ years of experience building and scaling production AI platforms and distributed systems. Deep expertise in LLM-powered systems, ML lifecycle management, and cloud-native infrastructure on AWS. Proven ability to design low-latency, highly available inference services, optimize performance and cost at scale, and collaborate across engineering, research, and product teams to deliver responsible, enterprise-grade AI systems.

Skills

AI & ML	LLM Inference, Agentic Systems, LangChain
	OpenAI APIs, NLP, Model Evaluation
ML Platforms	Feature Stores, Model Lifecycle, A/B Testing
	Monitoring, Experimentation
Cloud	AWS, Kubernetes, Docker, Terraform
	ECS, Lambda, API Gateway, CloudWatch
Distributed	High-Availability, Low-Latency Inference
	Event-Driven, Streaming, Scalability
Languages	Python, Go, TypeScript, C++, SQL
CI/CD	Jenkins, CircleCI, IaC, Automation

Nationality and Work Authorization

Citizen of the United States

Work Experience

JPMorgan Chase, Machine Learning Center of Excellence

Jersey City, New Jersey, USA

Lead ML Infrastructure Engineer

Nov 2023 - Current

- Spearheaded the design and implementation of a multi-region infrastructure for LLM Suite, ensuring high availability and scalable enterprise adoption.
- Led development of an agentic LLM framework for summarization and multi-step reasoning using LangChain and LangGraph.
- Designed and operated production LLM inference pipelines with observability, rate-limiting, and failure isolation to support enterprise-scale usage.
- Engineered Kubernetes-based inference services for chat-based AI assistants, optimizing resource utilization and system performance.
- Partnered with platform, product, and research teams to deliver secure, reliable AI capabilities across the organization.

Amazon Web Services, Glue

New York City, New York, USA

Software Development Engineer (Backend Infrastructure)

Sept 2022 - Aug 2023

- Led a team of 6 engineers to deliver AWS Glue support for large instance types, enabling high-memory, high-throughput ETL workloads.
- Designed and implemented backend solutions to mitigate hot partition issues, significantly improving service scalability.
- Eliminated recurring customer issues by introducing automated cleanup for leaked Elastic Network Interfaces (ENIs).
- Independently architected and drove features to reduce Glue job startup latency, improving customer time-to-insight.
- Participated in on-call rotations, triaging and resolving high-severity customer-facing production incidents.

Bloomberg LP

Princeton, NJ, USA

Senior Software Engineer

July 2020 - Sept 2022

- Led migration of critical financial data services from legacy C++ systems to event-driven, containerized Python microservices.
- Drove adoption of Kubernetes-based deployments, improving scalability, reliability, and operational consistency across production services.
- Designed distributed ingestion pipelines for high-volume financial and news data supporting downstream analytics and products.

Siemens Corporation
Princeton, NJ, USA

Machine Learning Engineer
Oct. 2017 - July 2020

- Principal Investigator and Technical Lead for a DARPA-funded project delivering scalable platforms for information extraction and document understanding.
- Architected and implemented end-to-end machine learning systems for complex data modalities, including point cloud datasets.

Bentley Systems Inc.
Watertown, CT, USA

Machine Learning Engineering Intern
Feb. 2015 - May 2015

- Developed and evaluated machine learning models for smart water networks to predict water usage and detect abnormal events in real time.
- Improved prediction accuracy by **2.2%** over a baseline ANN framework, reducing water leakage and false alarms on production datasets.
- Scaled machine learning pipelines using **AWS** and **GPU acceleration**, improving training and inference throughput.

University Information Technology Services, University of Connecticut
Storrs, CT, USA

Software Engineer, Part Time
Aug. 2015 - Jun. 2017

- Backend Developer in the UConn Facilities Asset Management Information System support team
- Integrated/Transitioned backend with Oracle, developed the data layer for the backend system.

Cognizant Technology Solutions
Chennai, TN, India

Software Engineer
Jan. 2009 - Jun. 2011

- Developed several features for claim processing engine for OptumInsight(client)
- Authored several database queries, stored procedures as part of database development.
- Actively participated in automating unit tests and integration tests.

Indian Statistical Institute
Kolkata, WB, India

Undergraduate Research Intership
Jan. 2008 - Aug. 2008

- Partial image encryption for secured multimedia communications

Education

University of Connecticut
Storrs, CT, USA

PhD, Computer Science and Engineering
Aug. 2011 - Jun. 2017

West Bengal University of Technology
Kolkata, WB, India

BS, Computer Science and Engineering
Aug. 2011 - Jun. 2017