Samir Sengupta

New York, USA • samir843301003@gmail.com • +1 551-359-1228
github.com/SamirSengupta • linkedin.com/in/samirsengupta • samir-sengupta.com

Professional Summary

Data Scientist and AI Engineer with 4+ years of experience developing enterprise-scale machine learning solutions, deep learning architectures, and production AI systems. Expert in end-to-end ML pipeline development, model optimization, and cloud deployment. Proven track record of delivering measurable business impact through advanced analytics, predictive modeling, and MLOps implementation. Specialized in computer vision, NLP, generative AI, and distributed computing. Currently pursuing Master's in Data Science. Authorized to work in the United States (CPT—OPT eligible).

Skills

Programming & Core Technologies: Python, SQL, R, Scala, Java, C++, JavaScript, Bash/Shell
ML/DL Frameworks: TensorFlow, PyTorch, Keras, XGBoost, LightGBM, CatBoost, scikit-learn, Hugging Face Transformers
Generative AI & LLMs: GPT, LLaMA, BERT, T5, Stable Diffusion, LoRA, QLoRA, PEFT, RAG, LangChain, OpenAI API
Computer Vision: OpenCV, YOLO, R-CNN, U-Net, GAN, VAE, Image Segmentation, Object Detection, OCR
Data Engineering: Apache Spark, Kafka, Airflow, Databricks, Snowflake, dbt, ETL/ELT pipelines, Data Lakes
Cloud Platforms: AWS (SageMaker, EC2, S3, Lambda, EMR), GCP (Vertex AI, BigQuery), Azure (ML Studio)
MLOps & DevOps: Docker, Kubernetes, MLflow, Kubeflow, CI/CD, Git, Jenkins, Terraform, Model Monitoring
Databases: PostgreSQL, MySQL, MongoDB, Redis, Elasticsearch, Neo4j, ClickHouse, Cassandra
Visualization & BI: Tableau, Power BI, Plotly, D3.js, Matplotlib, Seaborn, Streamlit, Dash
Statistical Analysis: A/B Testing, Hypothesis Testing, Bayesian Statistics, Time Series Analysis, Causal Inference

Professional Experience

AI Researcher (Nov 2024 – Present) — **Saint Peter's University**, New Jersey, USA

- Leading cutting-edge research in Large Language Models, implementing advanced fine-tuning techniques including LoRA and QLoRA for domain-specific applications, achieving 45% improvement in task-specific performance
- Architecting sophisticated RAG systems with vector databases (Pinecone, Weaviate) and semantic search capabilities, resulting in 60% improvement in contextual relevance and query accuracy
- Developing production-ready LLM solutions with custom tokenization and distributed training on multi-GPU clusters, reducing inference latency by 35%
- Mentoring 3 junior researchers and publishing findings in top-tier AI conferences

Software Developer (Aug 2023 – Jul 2024) — **Synradar**, Mumbai, India

- Architected and deployed real-time ML-based intrusion detection systems using ensemble methods and deep learning, achieving 40% improvement in threat detection accuracy and 25% reduction in false positives
- Implemented automated threat hunting platform using unsupervised learning and anomaly detection algorithms, processing 10M+ security events daily with 50% efficiency improvement
- Integrated LLaMA 3.1 with custom security knowledge base for intelligent code analysis and vulnerability assessment, improving code security evaluation by 30% and reducing manual review time by 60%
- Built interactive ML-powered security dashboards using React and D3.js, enabling real-time threat visualization and reducing incident response times by 35%
- Led cross-functional team of 5 engineers in implementing MLOps pipeline for continuous model deployment and monitoring

AI Solutions Architect (Jan 2022 – Jul 2023) — **Neural Thread**, Mumbai, India

- Designed and deployed end-to-end ML pipelines for predictive analytics, improving business forecast accuracy by 25% and generating $10K in additional revenue through optimized decision-making
- Developed state-of-the-art neural networks including Transformers and CNNs for multi-modal applications, achieving 30% performance improvement through advanced hyperparameter optimization and ensemble techniques
- Created production-grade generative AI solutions including conversational AI chatbots and content generation systems, increasing customer engagement by 35% and reducing support costs by 40%
- Optimized ML model performance for large-scale deployment using model compression, quantization, and distributed inference, reducing computational costs by 50% while maintaining 99.5% accuracy
- Implemented comprehensive A/B testing framework and statistical analysis for product features, leading to data-driven decisions that improved user retention by 20%

Education

Master of Science in Data Science (Sep 2024 – Dec 2025) — Saint Peter's University, New Jersey, USA
Advanced Coursework: Deep Reinforcement Learning, Advanced Neural Networks, MLOps, Distributed Systems, AI Ethics, Computer Vision, Advanced NLP, Causal Inference, Time Series Forecasting
GPA: 3.9/4.0 — Research Focus: Large Language Models and Multi-Modal AI Systems

Bachelor of Science in Data Science (Aug 2020 – Apr 2023) — University of Mumbai, Mumbai, India
Relevant Coursework: Advanced Statistics, Machine Learning Theory, Deep Learning, Data Structures & Algorithms, Database Systems, Linear Algebra, Probability Theory, Optimization Methods
Capstone Project: Multi-Agent Reinforcement Learning for Resource Allocation — GPA: 3.8/4.0

Key Projects & Research

IntelliBot: Enterprise-Grade Conversational AI Platform (Mar 2025)

- Developed custom LLM using LLaMA 3.3 architecture with domain-specific fine-tuning on 500K+ institutional documents
- Implemented advanced RAG with hierarchical document indexing and semantic chunking, achieving 85% accuracy in knowledge retrieval
- Deployed scalable microservices architecture on Kubernetes with auto-scaling capabilities handling 10K+ concurrent users
- Integrated advanced security features including PII detection, content filtering, and audit logging

AutoRecruit.ai: Intelligent Job Application Automation (Jan 2025)

- Built sophisticated NLP pipeline using transformer models for resume-job matching with 92% accuracy
- Implemented secure OAuth integration with LinkedIn API and intelligent form-filling using computer vision
- Created reinforcement learning agent for optimizing application success rates and personalizing cover letters
- Deployed full-stack application with React frontend and FastAPI backend, processing 1000+ applications daily

MedVision: AI-Powered Medical Diagnostic System (Dec 2024)

- Developed state-of-the-art computer vision models using Vision Transformers and ensemble CNNs for medical image analysis
- Achieved 95% accuracy in detecting anomalies across X-rays, MRIs, and CT scans using multi-modal deep learning
- Implemented explainable AI techniques (GRAD-CAM, SHAP) for model interpretability and clinical validation
- Created HIPAA-compliant web application with secure image processing and real-time diagnostic recommendations

TradingAlpha: Quantitative Trading Strategy Engine (Oct 2024)

- Built end-to-end algorithmic trading system using deep reinforcement learning and time series analysis
- Implemented real-time market data processing with Apache Kafka and low-latency model inference
- Achieved 23% annual return with 0.9 Sharpe ratio through advanced feature engineering and ensemble methods

Certifications & Professional Development

Advanced Certifications: TensorFlow Professional Developer — AWS Certified ML Specialty — Google Cloud Professional ML Engineer — Azure AI Engineer Associate — NVIDIA Deep Learning Institute — Stanford ML Specialization

Specialized Training: Advanced Deep Learning (DeepLearning.ai) — MLOps Engineering — Kubernetes Application Developer — Apache Spark Developer — HackerRank Certified Associate

Research & Publications

Peer-Reviewed Publications:

- "Efficient Fine-tuning of Large Language Models for Domain-Specific Applications" - Medium
- "Multi-Modal Fusion Networks for Enhanced Medical Image Analysis" - Saint Peters AI Conference 2025

Conference Presentations: 3 presentations at top-tier AI/ML conferences

Honors & awards

Technical Leadership: Led 15+ cross-functional ML projects — Mentored 8 junior data scientists and engineers — Established ML best practices and coding standards across teams

Business Impact: Generated $50K+ in cost savings and revenue through ML solutions — Reduced operational costs by 40% through automation — Improved customer satisfaction scores by 25% through personalization

Community Engagement: Active contributor to open-source ML projects — Technical blog with 5K+ monthly readers — Regular speaker at AI/ML meetups and conferences