Nainika Saha

Machine Learning Engineer

Location: San Francisco, CA | Mobile No: | @gmail.com | LinkedIn

Summary

Lead ML Engineer with 4+ years' experience building production-grade AI systems including RAG pipelines, generative NLP agents, and multimodal vision models. Proven track record of deploying scalable LLM infrastructure and driving cross-functional ML initiatives from experimentation to real-time deployment. Led a cross-functional initiative to develop and deploy a GPT-powered RAG platform using OpenAI, FAISS, and LangChain, enabling scalable document reasoning with multi-hop prompt chaining and self-consistency validation. Familiar with responsible AI principles including fairness, transparency, and model auditing for real-world deployment. Built predictive decision system to drive marketing optimization and targeting. Enabled 15% lift in customer retention through ML-driven targeting. Refactored and maintained core NLP pipelines and RAG infrastructure to support evolving GPT-based use cases, resolving critical deployment issues in a high-speed environment.

Skills

Programming and Scripting:	Python, SQL, Bash/Shell Scripting, Node.js, React, Java, Scala, C++, JavaScip
Mathematics and Statistics:	Linear Algebra, Calculus, Probability, Statistics, Optimization Techniques
Machine Learning and Data Science:	Supervised Learning, Unsupervised Learning, Reinforcement Learning, Deep Learning, NLP, Time-Series Analysis, Recommendation Systems
Frameworks and Libraries:	Scikit-learn, TensorFlow/Keras, PyTorch, XGBoost, LightGBM, Hugging Face
Data Engineering:	Apache Spark, Hadoop, Kafka, ETL Tools (NiFi, Talend, Informatica), dbt, Airflow, Dagster, DuckDB, LoRA, QLoRA, Tableau, Dash, Shiny
Data Wrangling & Visualization:	Data Cleaning/Preprocessing, Matplotlib, Seaborn, Plotly
Cloud and Deployment:	AWS (SageMaker), Azure (ML Studio), MLOps, Docker, Kubernetes, Flask, FastAPI
DevOps and Automation:	Git, Jenkins, CircleCI, GitHub Actions, Terraform, Ansible
Specialized Knowledge:	Computer Vision (OpenCV, YOLO), Anomaly Detection, AutoML (H2O.ai, AutoKeras), A/B Testing, feedback-loop retraining
Natural Language Processing:	NLTK, SpaCy, Hugging FaceTransformers, BERT, RNN, LSTM, Transform Models, PromptEngineering, Named Entity Recognition (NER)
Performance & Optimization:	Model Evaluation, Hyperparameter Tuning, Model Compression
Soft Skills:	Problem-Solving, Collaboration Tools (JIRA, Confluence), Communication
Generative AI:	GPT-4, LLaMA, Hugging Face Transformers, Stable Diffusion (familiar), LangChain, OpenAI APIs
Emerging Technologies:	Federated Learning, Edge Computing, ONNX, TFLite

Work Experience

KPMG, CA | Machine Learning Engineer **June 2024 – June 2025**

- Designed and deployed scalable ML models using Python, PySpark, and TensorFlow, reducing model training time by 40% through optimized multi-threading and load-balancing techniques.
- Led the development of a Natural Language Processing (NLP) pipeline using Transformers, Hugging Face, and OpenAI's GPT models, enhancing sentiment analysis accuracy by 30% for a financial services client.
- Implemented a high-performance data processing workflow on AWS, leveraging Vector DB and RAG Models, reducing data retrieval latency by 50% for real-time analytics. Developed and deployed a Retrieval-Augmented Generation (RAG) pipeline integrating OpenAI GPT-4 with FAISS and Pinecone for document-grounded question answering, achieving a 40% improvement in accuracy over baseline extractive models.
- Deployed ML models using MLOps best practices, integrating CI/CD pipelines, API Gateways, and automated model monitoring, improving model deployment efficiency by 60%.
- Optimized GPU-based deep learning models for fraud detection, increasing anomaly detection rates by 25% and reducing false positives by 20%, enhancing security for high-value transactions.
- Developed a custom feature engineering framework using PySpark and Scikit-learn, automating data preprocessing and improving model accuracy by 22% while reducing data pipeline execution time by 35%.
- Collaborated with cross-functional teams to build scalable APIs for AI service integration, supporting over 80,000 active users using tools like Flask, AWS, and Docker.
- Built internal analytics dashboards and reporting tools using Pandas, Plotly, and AWS QuickSight to support leadership and investor-facing insights.
- Assembled and prepared datasets for machine learning models, improving accuracy by up to 15% across Logistic

Regression, Naive Bayes, Decision Tree, Random Forest, and SVM using Scikit-learn.

Tech Mahindra, India | Machine Learning Engineer **June 2021 – July 2023**

- Designed and implemented scalable end-to-end pipelines for deploying machine learning models into production environments, utilizing Docker, Kubernetes, and cloud platforms such as AWS and GCP.
- Deployed a customer churn prediction model using TensorFlow and Flask, ensuring reliability, scalability, and high performance through the lifecycle. Implemented multi-hop reasoning with prompt chaining and self-consistency validation to improve factual correctness and answer robustness.
- Developed real-time monitoring systems using Prometheus and Grafana to track model accuracy, latency, and data drift, enabling quick identification and resolution of performance issues.
- Implemented CI/CD pipelines for automated deployment and versioning of machine learning models using Jenkins and GitHub Actions, reducing deployment time by 30%.
- Collaborated with data scientists, software engineers, and product teams to understand deployment needs and integrated models with existing data pipelines and RESTful APIs.
- Designed and maintained ML infrastructure on AWS, leveraging EC2, S3, and Lambda for efficient scaling and cost optimization.
- Ensured compliance with data privacy regulations such as GDPR and implemented role-based access controls to safeguard sensitive data during deployment.
- Conducted periodic performance evaluations and optimized deployed models using techniques like quantization and pruning, achieving a 20% reduction in latency.
- Integrated Kubeflow for model orchestration and monitoring, streamlining the ML operations lifecycle.
- Researched and adopted cutting-edge technologies such as MLflow, AirFlow, Dagster for tracking model experiments and deployment, improving operational efficiency by 25%.
- Designed a clustering-based customer segmentation pipeline using K-means and DBSCAN to identify high-conversion leads for targeted marketing campaigns.
- Collaborated with backend engineering teams to integrate ML outputs into customer-facing product features, enabling automated targeting at scale.

Education

Master of Science in Computer Engineering **Graduated, May 2025**
California State University Northridge, California, United States
Bachelor of Technology in Electronics and Communication Engineering **Graduated, May 2023**
Ramaiah Institute of Technology, Bengaluru, India

Projects

Autonomous Rover Control with Nasa's F Prime and YOLO On Raspberry Pi 5 GitHub | June 2024 – Mar 2025
Built and deployed real-time object detection model on Raspberry Pi 5 for in-field robotics inference; integrated STM32 for low-latency motor control and telemetry; used farm-like video feed simulations to train YOLOv8 model on structured fruit datasets. Used NVIDIA Jetson to train my ML model.
Autonomous Vehicle and Traffic Sign Detection Github | Sept 2024 – Dec 2024
Designed a YOLOv8-based traffic detection system achieving a mAP@50 of 0.63, deployed on Google Colab, and evaluated using FiftyOne.
Differential Privacy with Machine Learning and Deep Learning Sept 2022 – Feb 2023
Enhanced ML model privacy using AES-GCM with PATE and DP-SGD, ensuring 94% compliance on the MNIST dataset.

Certifications

- AWS Machine Learning, Coursera
- Natural Language Processing (NLP) and Chatbots, Coursera
- Introduction to Machine Learning: Language Processing, Coursera
- Introduction to AWS Identity and Access Management, Coursera