Aakarsh Sagar

AI & Machine Learning Engineer | LLM & MLOps Specialist
hello@aakarshsagar.com | +1 (720) 462-0865 | linkedin | github | Tokyo,Japan

PROFESSIONAL SUMMARY

Innovative and results-driven AI professional with a Master's in Data Science and a proven track record in designing, building, and deploying end-to-end machine learning systems. Specialized in Generative AI, LLMs, and MLOps, with deep expertise in NLP/NLU, Computer Vision, and scalable data pipelines. Passionate about solving complex problems by applying advanced techniques in Transformer architectures, fine-tuning, and RAG systems to deliver measurable business impact. Seeking remote roles to contribute to cutting-edge AI initiatives in the Japan market.

TECHNICAL SKILLS

- AI/ML Engineering: Transformer Models, LLMs (GPT-4, GPT-4o, HuggingFace), Fine-tuning, RAG, Generative AI, Predictive Modeling, Deep Learning, Anomaly Detection, Clustering
- Programming & MLOps: Python, SQL, PySpark, R, MLflow, Docker, Git, CI/CD, FAISS, Vector Databases, Model Evaluation, Reproducibility
- Frameworks & Libraries: PyTorch, TensorFlow, OpenCV, YOLOv8, ResNet, Scikit-learn, Pandas, NumPy, NLTK, spaCy
- Data & Cloud Engineering: ETL/ELT Pipelines, Data Quality, Feature Engineering, Automation, AWS (SageMaker, S3), GCP, Azure, Databricks, Big Data
- Visualization & Deployment: Streamlit, Power BI, Tableau, REST APIs, Model Deployment, Monitoring

WORK EXPERIENCE

AI Engineer (NLP/LLM Focus)
Saayam For All | *San Jose, CA (Remote)* | ***Oct 2024 – Present***

- Architected and productionized a multilingual speech-to-text ETL pipeline using Meta's SeamlessM4T model via HuggingFace, processing voice notes across 15 languages for a global user base.

- Engineered a robust data preprocessing and feature engineering workflow to clean noisy audio and extract key acoustic features (pitch, tone, tokens), improving input data quality for model inference.
- Established a continuous model evaluation framework using Character Error Rate (CER) and Precision/Recall, achieving an 80% average transcription accuracy - a 14% improvement over the baseline.
- Applied unsupervised learning (clustering) and regression analysis on operational data to identify key inefficiencies, providing data-driven recommendations that improved system throughput by 15%.
- Collaborated with engineering and product teams to integrate performance insights into a live Power BI dashboard, enabling real-time monitoring and stakeholder decision-making.

Machine Learning Engineer (MLOps & Research)
University of Denver | *Denver, CO* | ***Jul 2023 – Apr 2024***

- Designed and deployed a production-grade RAG system using GPT-4o and a FAISS vector database, incorporating synthetic data generation for improved retrieval, which reduced average IT ticket resolution time by 25 minutes.
- Built a full lifecycle evaluation framework measuring accuracy, recall, coverage, and diversity, creating benchmarks for iterative model improvement and ensuring reproducibility.
- Curated and annotated a proprietary ServiceNow dataset; implemented a redundancy detection system that cut duplicate tickets by 30% and boosted retrieval recall by 15%.
- Automated the large-scale migration of user portfolios from on-premise to Azure Cloud using Power Automate, reducing manual transition time by 33%.
- Developed interactive Tableau dashboards to forecast usage trends and recommend resource allocation, driving a 60% increase in platform adoption.

Data Scientist (Big Data & Analytics)
Infosys | *Springdale, AR (Remote)* | ***Sep 2019 - Jun 2022***

- Engineered and optimized automated, scalable data pipelines using SQL and PySpark on AWS and GCP infrastructure, supporting business-critical data operations and improving data reliability by 25%.
- Performed advanced time-series analysis and forecasting on performance data for 800+ distributed systems, optimizing resource allocation and resulting in a 10% increase in system reliability.
- Built predictive analytics dashboards to track KPIs and performance trends across manufacturing plants, enabling data-driven process optimization.

- Identified early failure patterns by applying anomaly detection algorithms to system backup logs, enabling proactive maintenance across 3 business units.

MACHINE LEARNING PROJECTS

Generative AI E-Commerce Support Assistant
Python, GPT-4o-mini, FAISS, Streamlit | ***May 2025 – Jun 2025***

- Developed an end-to-end RAG pipeline for a customer support AI assistant, involving dataset synthesis, curation, and benchmarking with BLEU/ROUGE metrics.
- Integrated FAISS for efficient vector retrieval with an LLM generator fine-tuned for brand-appropriate tone, significantly reducing manual ticket handling.
- Deployed a containerized application with automated regression tests to ensure reproducibility and production reliability.

Computer Vision: Real-Time Object Detection System
Python, PyTorch, YOLOv8 | ***Mar 2024 – Jun 2024***

- Conducted exploratory data analysis (EDA) on bounding box annotations, uncovering critical labeling errors that prompted a new model development strategy.
- Fine-tuned a YOLOv8 model on a custom dataset, achieving state-of-the-art performance with 0.99 Recall and 0.994 mAP.
- Optimized and delivered a production-ready model that boosted real-time detection accuracy and reduced missed objects by over 95%.

EDUCATION

University of Denver, Denver, CO
Master of Science in Data Science | GPA: 4.0/4.0 | ***Sep 2022 – Aug 2024***

New Horizon College of Engineering, Bangalore, India
Bachelor of Engineering in Electrical Engineering | GPA: 3.42/4.0 | ***Aug 2015 – Jun 2019***

(Open to Remote - Japan Time Zone)