SANKET MUCHHALA

AI /ML Engineer

+1 812-778-4451 | sanketmuchhala1@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI/ML Engineer with 3+ years of experience building scalable solutions using generative AI, LLMs, and NLP. Skilled in designing agentic systems, document intelligence workflows, and ML pipelines deployed at scale. Adept at driving automation and data-backed insights across industries like insurance, esports, and enterprise analytics.

EDUCATION

Master of Science in Data Science| Indiana University Bloomington, USA| Aug 2022 – May 2024

Bachelor of Technology in Information Technology| Thakur College of Engineering and Technology, India| Aug 2018 – May 2022

TECHNICAL SKILLS

Programming Languages:	Python, SQL, R, JavaScript
AI/ML Frameworks & Tools:	Scikit-learn, TensorFlow, PyTorch, FastAPI, MLflow, SpaCy
Generative AI & LLMs:	GPT-4 (via OpenAI APIs), LangChain, RAG, Agentic AI tools, Vector DBs(FAISS)
NLP & Document Intelligence:	Named entity recognition (NER), text classification, summarization, sentiment analysis
Data Engineering & Storage:	Pandas, NumPy, PySpark, AWS (SQS, Step Functions), Azure Data Lake, Azure SQL
Databases:	PostgreSQL, MySQL, Snowflake, Teradata
Visualization & BI:	Tableau, Power BI, R Shiny

PROFESSIONAL EXPERIENCE

AI Engineer| Progressive Insurance, Remote, USA | May 2024 – Present

Project: AI-Powered Claims Automation & Risk Analysis (ACARA)

- Engineered custom NLP and CV models using TensorFlow and PyTorch to process and classify claim-related texts, forms, and images.
- Developed NER, sentiment analysis tools using BERT models via Hugging Face Transformers to extract key entities from claim documents.
- Built ML pipelines with Apache Airflow, Azure Data Factory to preprocess and stream data from on-prem SQL Server and Azure Data Lake.
- Deployed models to production using Azure ML Services and managed versioning with MLflow and DVC.
- Designed RESTful APIs using FastAPI to integrate predictive models into the core claims processing system secured endpoints using OAuth2.
- Containerized model services with Docker and orchestrated deployment via Kubernetes on AKS.
- Established observability infrastructure using Prometheus and Grafana, with alerts and monitoring integrated into Azure dashboards.
- Achieved 35% reduction in manual claim processing time and improved fraud detection accuracy by 25%.
- Collaborated with underwriters and legal compliance teams to ensure explainability and auditability of ML models per regulatory standards.

Research Assistant – Generative AI | Indiana University Bloomington, IN, USA | Dec 2023 – May 2024

- Improved transcript accuracy by 18pp using a GPT-4 RAG pipeline deployed on BigRed200, processing over 200 hours of esports videos.
- Reduced latency 40% in chat feature via GPT-4 sentiment analysis microservice, processing 1M+ messages in near real-time.
- Automated retraining pipelines using SLURM on HPC systems, cutting manual ETL effort by 6 hours per match.
- Documented GenAI workflows, adopted by two graduate cohorts for ongoing esports psychology research.

Data Analyst | IBM, MH, India | Sep 2020 – Jun 2022

- Led end-to-end development of a churn prediction model using Python and Scikit-learn, driving a 20% reduction in customer attrition.
- Refactored ETL workflows using Azure Data Lake and SQL, improving data availability and cutting processing time by 15%.
- Built automated data validation pipelines with SQL and Python, raising dashboard reporting accuracy by 18%.
- Deployed ML models to Azure ML environments with CI/CD support, accelerating release cycles by 25% across internal products.
- Authored reproducible model documentation and Jupyter-based reports to align analytics delivery with stakeholder needs.
- Partnered with business analysts and solution architects to prioritize model features based on customer behavior data trends.
- Introduced versioning standards for ML pipelines and datasets, increasing transparency in model updates and audits.
- Conducted internal training sessions on Python-based analytics tooling, improving team adoption of reusable code modules.

PROJECTS

RAG Search Engine: GitHub

- Built a Retrieval-Augmented Generation (RAG) system using GPT-3 and LangChain for multi-source question answering.
- Integrated FAISS-based vector search with PDF chunking to enable fast, contextually relevant retrieval.

Job Description Keyword Extractor: GItHub

- Created an NLP-based web tool (SpaCy, LangChain) to extract role-specific keywords from job descriptions.
- Improved resume-to-JD matching accuracy by 15% across 500+ users.

Location-Based File Sharing System: Live Demo

- Engineered a serverless AWS solution using S3 and Lambda with geospatial filtering, maintaining 99% uptime with efficient access control.
- Integrated OpenStreetMap via Leaflet.js into a responsive JavaScript frontend, enabling real-time file discovery within user-defined radii.

Certifications:

AWS Certified MLE Associate, Azure AI Fundamentals(AI-900), Databricks Generative AI Foundations, Google Data Analytics