Taha Ali

Lead GenAI Engineer | Senior Machine Learning Engineer | Expert In AI, ML, GenAI, LLM's and Data Science

tahanur840@gmail.com (612) 208-3668 Minneapolis, Minnesota,United States

Profile

Proficient Data Scientist with expertise in Python, TensorFlow, PyTorch and deep learning (CNNs, LSTMs, Transformers); specialized in NLP & GenAI (LLMs, RAG, on-prem LLMs, chatbots); skilled in data engineering (ETL, SQL, Airflow), RPA automation and MLOps/DevOps pipelines; experienced in healthcare analytics and scalable ML deployments on GCP, AWS & Azure; delivered end-to-end AI solutions.

Professional Experience

09/2021 – present

Ecologix

AI/ML Engineer | Conversational AI, Generative Systems, and Scalable LLM Solutions

- Developed domain-specific AI calling agents and agentic models, optimizing task performance in niche sectors through hybrid approaches that balanced compute efficiency with high model accuracy.
- Designed and implemented scalable production-grade solutions for integrating LLMs, focusing on cost-effective deployment, latency reduction, and inference optimization.
- Engineered a full-stack Retrieval-Augmented Generation (RAG) system using OpenAI, Anthropic Claude, and Google Gemini Vertex AI, orchestrated with LangChain and monitored via LangSmith.
- Built real-time, WebSocket-enabled conversational AI bots using FastAPI and integrated backend services for dynamic RAG interactions.
- Automated embedding workflows with Cloudflare Workers and VoyageAI, and architected a robust text embedding pipeline using OpenAI Text-Embedding-002, migrating vector storage from MongoDB Atlas to Elasticsearch and Pinecone for efficient semantic search.
- Led intelligent agent development with CrewAI and deployed microservices on Google Cloud Run with CI/CD pipelines built using GitHub Actions, Cloud Build, and Cloud Run.
- Implemented headless data scraping solutions on DigitalOcean VMs using Scrapy and Selenium, enabling structured data extraction from complex childcare platforms.
- Delivered computer vision projects, including sports object tracking and re-identification using Siamese networks, CLIP, Vision Transformer, and custom detection models.
- Conducted swimming pose estimation using YOLO Pose and OpenCV, enabling performance analytics through stroke detection.
- Built a proof-of-concept for automated video content insights using the Twelve Labs API.
- Automated healthcare billing and medical coding workflows using RPA tools, enhancing operational efficiency in clinical documentation and insurance claims.

- Utilized LLaMA 3.2 visual models for OCR to extract detailed patient claim data from scanned documents, improving accuracy in healthcare automation.

05/2016 – 08/2021 **Estela**

AI/ML Engineer | Computer Vision, Speech Processing, and Scalable ML Systems

- Trained and fine-tuned advanced computer vision models including YOLOv5, Ultralytics, and InceptionV2, leveraging VGG16 for deep feature extraction to maximize detection accuracy across real-time applications.
- Built fully automated ML pipelines for data ingestion, model training, validation, and testing, enabling rapid iteration and reproducibility across computer vision and NLP projects.
- Developed scalable RESTful APIs for model serving using FastAPI and Pydantic, containerized with Docker, backed by PostgreSQL, and deployed on AWS EC2 with CI/CD pipelines using GitHub Actions and Jenkins.
- Orchestrated distributed microservices and background tasks using Celery, Redis, RabbitMQ, and WebSocket for real-time data streams and notifications.
- Applied algorithmic optimizations and GPU acceleration to enhance real-time analytics throughput (from 5 fps to 33 fps).
- Managed data pipelines using shell scripting, AWK, and Apache Airflow; performed data augmentation and ETL operations to increase model robustness and reliability.
- Engineered NLP classifiers using TF-IDF, Random Forest, and Doc2Vec embeddings; implemented model stacking, detailed performance metrics, and explainability reports using SHAP and LIME.
- Implemented model optimization techniques such as quantization, pruning, and mixed-precision inference to reduce deployment latency and resource consumption.
- Leveraged Kubernetes and Docker Compose for container orchestration, supporting auto-scaling, high availability, and rolling updates in production.
- Set up monitoring and alerting infrastructure using Prometheus, Grafana, and AWS CloudWatch to track system health and detect anomalies.
- Led the architecture and deployment of state-of-the-art speaker recognition and speaker separation systems, including enrollment pipelines integrated with secure user databases.
- Designed real-time speech enhancement and denoising algorithms for streaming and offline environments, improving clarity in noisy conditions.
- Developed a private information masking system to redact sensitive data from call audio, ensuring compliance with data privacy standards.
- Conducted research and implementation across key voice domains including Text-to-Speech (TTS), Speech-to-Text (STT), speaker separation, and speech privacy.
- Spearheaded the development of high-fidelity Korean TTS models to elevate voice synthesis quality for multilingual applications.
- Promoted engineering best practices through rigorous unit/integration testing, code reviews, MLflow tracking, and comprehensive documentation.

Skills

- Artificial Intelligence
- Deep Learning
- Data Analysis & Visualization
- Rotating Proxies
- Sockets
- Keras
- PyCharm
- Numpy
- C/C++
- Docker
- Automated Machine Learning (AutoML)
- Natural Langugae Processing
- Data Gathering
- Python
- Scikit-learn
- TensorFlow
- MLOps
- LSTM
- Time Series Analysis
- Pandas
- ElasticSearch
- RESTful APIs
- Computer Vision
- Scrapy
- Selenium Data Scraping
- Django
- TensorRT
- Image Processing
- Fast API
- Shell & AWK
- Hadoop
- CI/CD Pipelines

Projects

RailVision AI

- Spearheaded development of a smart railway monitoring system utilizing LiDAR point clouds and 2D imagery for locomotive detection and rail track segmentation.
- Engineered algorithms for locomotive tracking, turnaround point detection, track partitioning, and automated alert generation.
- Focused on AI-driven computer vision using Python.
- Trained a 3D recognition model for locomotive identification from LiDAR-generated point clouds.
- Designed a 2D vision model for identifying locomotives in imagery.
- Experienced with KITTI dataset for autonomous navigation and scene understanding.
- Directed a team of annotators, established workflows for labeling, and developed annotation/training pipelines to streamline data operations.

StreamDetect

- Built a live vehicle monitoring system using image processing techniques for classifying real-time traffic from surveillance feeds.
- Developed classification logic to distinguish five specific vehicle types including public transport variants.
- Integrated vehicle counting and temporal analytics to identify traffic trends during peak hours.
- Designed and deployed the solution using Django and AWS cloud infrastructure.
- Used Python for object detection, image classification, and back-end development.

WebMiner Pro

- Worked with a leading data extraction company to automate large-scale web scraping tasks for enterprise clients.
- Used a robust Python-based stack including Scrapy, BeautifulSoup, and regex for frequent and resilient web data extraction.
- Developed techniques for circumventing anti-bot measures such as CAPTCHA and IP bans, using rotating proxies and internal APIs.
- Transformed scraped data into structured formats for AI-driven analytics.
- Created a Selenium-based tool to collect and analyze job listings from LinkedIn.
- Applied NLP techniques for extracting metadata like job seniority and salary using question-answering models.

NordicInvoiceIQ

- Designed an invoice processing solution for Danish hotel vendors involving OCR and Named Entity Recognition (NER).
- Tackled irregular document layouts in PDF/image formats through a custom vision pipeline.
- Employed YOLOv8 to detect key data zones on invoices, followed by OCR text extraction.
- Fine-tuned a BERT model to perform domain-specific NER on extracted invoice data.
- Architected the system as a Flask API for scalability and ease of deployment.
- Led a 5-member team and maintained direct communication with clients for agile project iteration.

RetailVision

- Developed a computer vision-based retail analytics tool for product detection on store shelves.
- Architected a scalable microservices solution integrating various ML components.
- Implemented image stitching to produce panoramic views of shelves for enhanced analysis.
- Utilized Minio for managing ML assets and Docker for consistent deployment.
- Created FastAPI endpoints to expose trained model predictions via RESTful APIs.
- Used PostgreSQL for persistent data storage and retrieval within the application.

InfluenceGraph

- Built a data intelligence platform to analyze and categorize online creators based on social media content.
- Designed recommendation algorithms to map content creators to thematic communities.
- Developed a classifier to determine creators' core niches using content metadata.
- Contributed to the design and development of a visual dashboard for real-time data insights.
- Utilized GPT-3 for enriching Instagram profiles with generated biographies.
- Applied NLP and computer vision techniques in Python.
- Incorporated ElasticSearch and Kibana for efficient search and data visualization.

Education

Master in Computer Science
San Diego State University