USAMA YASEEN

+92 348-6622291 ⋄ Lahore, Pakistan

usamayaseen867@gmail.com ⋄ linkedin.com/in/usamayaseen867

SUMMARY

I am an experienced Data Scientist with 2 years of hands-on work in machine learning, deep learning, and large language models (LLMs). I specialize in building intelligent systems using Python, advanced NLP, and cutting-edge AI frameworks like LangChain, LangGraph, and LangFlow. My work includes training and fine-tuning deep learning models such as Transformers, RNNs, LSTMs, CNNs, and YOLO. I have also developed multi-agent systems for both natural language processing and computer vision tasks, making complex workflows more efficient and intelligent. A big part of my focus is on deploying scalable Retrieval-Augmented Generation (RAG) systems and real-world NLP solutions with a strong emphasis on MLOps. I am skilled in cloud deployment and have experience with big data tools like Hadoop and Apache Spark to handle and process large-scale data. Above all, I enjoy turning complex technical challenges into practical, data-driven solutions that create real impact.

EDUCATION

BS Computer Science

University of Central Punjab   2020 – 2024

- Relevant Coursework: Calculus, Linear Algebra, Statistics, Numerical Computing, Introduction to NLP, Intro to ML, Intro to Deep Learning, Parallel and Distributed Computing.

WORK EXPERIENCE

Lead Data Scientist

Aug AI   Feb 2025 – Present

- Leading a team in developing AI solutions that enhance human intelligence.
- Overseeing projects involving LLMs, RAG systems, and multi-agent architectures.
- Mentoring junior data scientists and fostering a collaborative environment.usamayaseen-datascience.github.io

Data Scientist

Aug AI   Mar 2024 – Feb 2025

- Developed and deployed AI models for various applications, including medical diagnostics and financial analysis.
- Collaborated with cross-functional teams to integrate AI solutions into existing systems.

Machine Learning Engineer (Trainee)

Aug AI   Dec 2023 – Mar 2024

- Assisted in developing machine learning models and data pipelines.
- Gained hands-on experience with NLP frameworks and cloud-based AI deployments.

Machine Learning Engineer

SESQUAD   Jul 2023 – Dec 2023

- Worked on deep learning projects, including medical imaging and chatbot development.
- Implemented AI solutions using frameworks like TensorFlow and PyTorch.

PROJECTS

AI-Powered Research Assistant

- Built an intelligent research assistant that automates research, scraping, and report generation.
- Handled natural language queries to deliver summarized and well-formatted content.
- Used BeautifulSoup, Scrapy, and Selenium for deep web data extraction.
- Generated downloadable reports in PDF and Word formats with PDFKit and python-docx.
- Integrated GPT-3/4 to enhance clarity and coherence of responses.
- Designed the backend using FastAPI for smooth query processing and automation.
- Significantly reduced manual research time and improved content generation quality.

Automated Data Analyst Agent with Graph Generation

- Developed a tool to automate data analysis and graph creation from uploaded CSV files.
- Used Pandas and Matplotlib to visualize data and generate meaningful insights.
- Provided natural language explanations using GPT-3/4 integrated via LangChain.
- Enabled users to explore trends and patterns without manual coding.
- Simplified data storytelling for non-technical stakeholders.

Financial Bot for Valuation and Earnings Call Analysis

- Designed a smart financial assistant capable of analyzing valuations and earnings calls.
- Used MongoDB and vector databases for efficient retrieval of financial documents.
- Integrated GPT-4 mini for summarizing transcripts and extracting financial insights.
- Built a secure API with JWT auth, enabling multiple concurrent financial bots.
- Employed an agentic architecture to delegate tasks across specialized agents.
- Improved the speed and accuracy of financial research and decision-making.

Quantum-Infused LLM-Based Knowledge Retrieval

- A research project focused on addressing real-time and domain-specific shortcomings of LLMs through a smart, multi-model pipeline.
- Identified the core challenge of LLMs falling short in specialized, real-time knowledge tasks.
- Led the design of a multi-model pipeline combining feature engineering, advanced scraping, and summarization techniques.
- Collaborated closely with faculty and research supervisors to refine methodology and ensure robust implementation.

Legal Assistant – Multi-Agent LLM Chat for Legal Documents

- Built for legal consultants to navigate and analyze bulk legal documents with conversational AI.
- Spearheaded the development of multiple RAG modules using GPT-4o, GPT o1-preview, Claude Sonnet, Claude Opus, and Gemini Flash Pro 1.5.
- Integrated Weaviate as the core vector database across the entire MVP.
- Worked on voice-based interaction using Deepgram and AssemblyAI for live speech transcription and document conversion.

Calling Agent using LiveKit

- A real-time calling agent to handle calls and filter spam via live transcription analysis.

- Integrated LiveKit to handle inbound and outbound calls.

- Developed real-time transcription pipelines and spam detection logic.

- Designed and maintained caller schema for managing whitelist and blacklist functionality.

Financial Assistance AI

- Created a system for personal financial forecasting and advisory.

- Trained ML models to predict future expenses based on user behavior.

- Developed a chatbot using fine-tuned LLMs and Retrieval-Augmented Generation (RAG).

- Used historical data for personalized and context-aware financial suggestions.

- Helped users improve financial habits over a 6-month period.

- Built with Python, LangChain, TensorFlow, and AWS SageMaker.

Fraud Detection System for Banking Transactions

- Designed an end-to-end fraud detection platform for real-time banking data.

- Applied anomaly detection to flag unusual transaction patterns with high accuracy.

- Deployed the system using AWS SageMaker, S3, and CI/CD pipelines.

- Integrated explainable AI using Anthropic's Claude API to justify fraud decisions.

- Significantly reduced false positives and enhanced fraud prevention measures.

- Built a scalable, secure architecture using Python and AWS cloud services.

Bikes on Lifters Detection – Atlas Honda

- A smart detection system to automate bike counting in Honda workshops, improving efficiency tracking.

- Collected and annotated 8500+ frames from real workshops; fine-tuned YOLOv11n for accurate bike detection.

- Integrated the model into Atlas Honda's service portal for real-time usage.

- Deployed solution on live servers with automation scripts for seamless updates.

Facial Recognition-Based System

- A full-stack face recognition system for secure identity verification using deep learning models.

- Led backend development with Django and PostgreSQL for embedding storage and comparison.

- Used YOLOv8s for accurate face detection and FaceNet for embedding extraction.

- Trained a custom Inception V3 model to distinguish between human and non-human faces.

Cluster Merging (Inspired by HDBSCAN)

- Developed a custom clustering solution to merge clusters based on density metrics.

- Researched and built a novel approach to mimic HDBSCAN-level clustering without full dependency.

- Implemented density analysis to merge meaningful clusters and filter out noise.

- Successfully handled large, pre-labeled datasets and improved cluster coherence.

Heat Exchanger Data Prediction

- Predicted complex time-series data from heat exchangers using machine learning.
- Tested models including LSTM, Transformers, XGBoost, and Random Forest.
- Performed thorough evaluation and selected Random Forest with 500 trees as the best model (RMSE: 0.02).
- Led the model training, tuning, and comparative analysis to ensure reliability.

SKILLS

Machine Learning & Deep Learning:	Proficient in TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, LightGBM, CatBoost. Skilled in CNNs, RNNs, Transformers, Vision Transformers, and transfer learning techniques.
Natural Language Processing (NLP):	Experience with Hugging Face Transformers, BERT, GPT models, T5, RoBERTa, BART, GloVe, NLTK, Spacy, and advanced RAG systems using LangChain, LangGraph, and LlamaIndex.
Data Engineering & Processing:	Strong in data handling with Pandas, NumPy, Apache Spark, and Airflow. Skilled in data annotation, model evaluation (GridSearchCV, Optuna), and performance metrics (Precision, Recall, ROC AUC).
Statistical Analysis:	Knowledge of Bayesian inference, hypothesis testing, A/B testing, and statistical modeling.
Computer Vision:	Worked with YOLO (v5s, v8s, v11n, v11s), FaceNet, InceptionV3, MediaPipe, and OpenCV for image processing, face recognition, and pose estimation.
Vector Databases & RAG:	Hands-on experience with FAISS, Pinecone, Weaviate, ChromaDB, Milvus, and integration into RAG pipelines for semantic search.
Generative AI & Assistant APIs:	Integrated GenAI tools using LangChain, OpenAI APIs (GPT-4/4o), LlamaIndex, and Assistant APIs for multi-agent systems and domain-specific applications.
Backend & Development:	Skilled in Django, Flask, FastAPI, Celery, MySQL, MongoDB, Docker, and REST API development.
Programming Languages:	Proficient in Python, C, C++, Java, and JavaScript.
Cloud & Big Data Tools:	Experience with AWS (SageMaker, Lambda, S3, EC2), Hadoop, and Apache Kafka for scalable data processing.
Visualization & Reporting:	Comfortable with Matplotlib, Seaborn, Plotly, and Tableau for presenting insights and building dashboards.
Version Control & Collaboration:	Git, GitHub, GitLab, Bitbucket.

CERTIFICATIONS

- ChatGPT Prompt Engineering for Developers
- Building Multimodal Search and RAG
- Intro to Deep Learning
- Computer Vision
- Deep Learning Specialization