YASH PAREEK

github.com/pareek-ml yash.pareek@usi.ch pareek-ml.github.io linkedin.com/in/pareek-yash
+41 (076) 244 15 08

Dynamic ML engineer with proven skills in LLM fine-tuning and deployment—transforming enterprises with RAG applications, real-time forecasting systems, and computer vision solutions delivering tangible business impact.

Professional Experience

Machine Learning Engineer January 2023 – August 2024
Perimattic Bengaluru, IN

- Deployed a RAG-LLM chatbot across 10 airports, achieving 80% response accuracy and 70% faster replies, by fine-tuning LLAMA 2 on enterprise data.
- Enabled scalable, secure deployment of 15GB+ enterprise datasets by building an on-prem LLM with biometric authentication using Kubernetes, Docker, and MLflow.
- Delivered real-time hourly forecasting for airport parking using a full ML pipeline with NeuralProphet and TimescaleDB.

Freelance ML Developer March 2022 – July 2023
Jaipur, IN

- Reduced vehicle in-out time and automated charge calculation by deploying a YOLO-based License Plate Recognition system on AWS, achieving a 30% reduction in inference time.
- Boosted analysis efficiency with up to 5x content compression by deploying transformer-based models to summarize daily news and sales calls.
- Increased monthly sales by 30% by uncovering customer preferences through sentiment analysis of Amazon reviews, which informed pricing strategy.

Projects

RAG Application for SQL Query Generation | Built an end-to-end RAG pipeline that converts natural language queries into SQL. The system identifies the best-fitting schema and generates the corresponding SQL query. | *LLAMA 3.2, FastAPI, Docker, ChromaDB, TypeScript*

Breast Cancer Detection (Wisconsin) | Developed a complete ML pipeline to classify breast cancer tumors (benign vs. malignant) using feature selection, outlier detection, and statistical testing. Achieved 95.6% accuracy with perfect precision. | *Random Forest, Hydra, Scikit-Learn, Docker*

Voice-Driven LLM Chat Interface | Architected a real-time, hands-free conversational system using fine-tuned Whisper for speech-to-text and Coqui TTS for text-to-speech synthesis. | *Python, OpenAI, Streamlit*

LLM Fine-Tuning with QLORA | Enhanced Vicuna and LLaMA-13B performance by applying QLORA and PEFT on curated datasets, yielding more accurate and context-aware responses.

Education

USI Lugano, Masters of Science in AI September 2024 – Present
NLP, Machine Learning, Deep Learning, Data Design & Modelling Lugano, Switzerland

Rajasthan Technical University, B.Tech. Computer Science August 2018 – July 2022
Programming, Machine Learning, Algorithms, Big Data Graduated with Distinction

Skills

PyTorch | TensorFlow | HuggingFace | NLP | Python | NumPy | Pandas | Statistical Analysis
MLOps | CI/CD | DVC | Docker | Git | AWS ML | Azure | PySpark | OpenAI
English: Full Professional Fluency
Residence Permit B: Full time allowed