Karan Singh Chauhan
6058 Rothko Ln, Simi Valley CA | +18125388341 | Karan.m0624@gmail.com | linkedin.com/in/karan-chauhan-49219213 |https://github.com/karan0624

Professional Summary

Data Scientist and Machine Learning Engineer with proven expertise in Generative AI, predictive modeling, and scalable data systems. Skilled in designing and deploying LLM inference pipelines using vLLM, TGI, RAG, and vector databases, applying advanced optimizations like 4-bit quantization (AWQ) and speculative decoding for efficient model performance. Experienced in statistical learning, forecasting, and multimodal AI, delivering real-world impact across healthcare analytics, computer vision, and enterprise automation. Strong in applied research and experimentation, bridging theoretical innovation with production-ready, high-performance AI solutions that drive measurable business outcomes.

Technical Skills

- **Inference Optimization: vLLM, TGI, ONNX Runtime, quantization (4-bit, AWQ, GPTQ), speculative decoding**
- **Low Latency Systems: Caching, batching, tensor parallelism, distributed inference**
- **LLM Deployment: Containerized serving (Docker, Kubernetes), FastAPI, scalable REST APIs**
- **Vector & RAG Pipelines: FAISS, Pinecone, Weaviate, Chroma for retrieval-augmented LLMs**
- **Machine Learning: PyTorch, Hugging Face Transformers, Scikit-learn**
- **Cloud & Infrastructure: Azure, AWS, Kubernetes, CI/CD (GitHub Actions)**
- **Programming: Python, SQL, CUDA (beginner), Triton (exploring)**

Experience

Research Assistant | University of Alabama at Birmingham, Birmingham, AL **09/2025 – Present**

- Conducting **research** on **AI, LLMs, NLP**, and **Bioinformatics,** applying advanced machine learning to healthcare and computational biology problems.
- Developing and evaluating models with **PyTorch, TensorFlow**, and **Hugging Face Transformers** for **biomedical text analysis** and data interpretation.
- Performing **data preprocessing, feature engineering, and statistical analysis** on **clinical** and **genomic datasets** to support experimental studies.

CETS AV/IT Technician | Indiana State University, IN, US **02/2024 – 05/2025**

- **Configured and managed campus IT & network infrastructure** using Cisco switches, system imaging, domain joins, and static IP mapping to onboard and maintain devices across the college.
- **Installed, maintained, and troubleshot AV systems** (Extron, projectors, HDMI switching), ensuring seamless classroom delivery and zero downtime for campus-wide setups.

Data Scientist – Industrial / Manufacturing Analytics | Make Brass Industries, GJ, India **01/2023 – 05/2023**

- **Built and deployed real-time ML systems** (YOLOv5 + OpenCV on Raspberry Pi) for brass manufacturing, reducing surface defect rates by 25% and minimizing manual inspection.
- **Developed predictive models** (XGBoost, time-series forecasting) to optimize export forecasting and inventory planning, improving delivery schedule accuracy.
- **Engineered automated data pipelines** with Azure Data Factory + SQL, integrating IoT sensor data for real-time tracking and presented insights to leadership in client-facing settings.

Machine Learning Engineer | Limelight IT & Research Pvt. Ltd., Ahmedabad, GJ, India **04/2021 – 06/2022**

- **Reduced ML training and inference time by 45%** and improved data reliability by 35% by developing real-time AI pipelines with Databricks, Apache Spark, and anomaly detection models (LSTM, XGBoost).
- **Deployed scalable containerized microservices** by integrating FastAPI with Kubernetes and Docker, enabling RAG-style health telemetry inferences.
- **Improved support response efficiency by 20%** by automating diagnostics and ticket triaging through Python and PowerShell scripts.

Projects

Role of Agentic AI in Eradication for the Need of Scrum Master (FastAPI, Docker, LangChain, GPT-4, Pinecone/FAISS, Slack API, Jira API, GitHub API) Self-initiated research & development (2024 – Present)

- **Developed an AI Scrum Master Assistant** that converts live meeting speech into structured sprint updates and Jira tasks using speech-to-text, LLaMA 3, and RAG pipelines.
- **Optimized multimodal inference pipelines** with vLLM (token streaming, speculative decoding) and vector databases (FAISS/Weaviate), boosting contextual accuracy and reducing latency.
- **Deployed cost-efficient large-scale models** with 4-bit quantization, Docker, and Kubernetes, delivering scalable real-time performance for actionable team intelligence.

Speech Emotion Recognition - (Deep Learning, PyTorch, CNNs, Streamlit)

- **Achieved 92% accuracy** by developing a transformer-based speech emotion recognition model (30M parameters) with ResNet-based audio embeddings.
- **Improved training stability by 40%** through adaptive gradient clipping and batch normalization.
- **Enhanced preprocessing and feature engineering workflows** using Python libraries (pandas, NumPy) for audio data exploration.

Diabetic Readmission Prediction - (TensorFlow, FastAPI, Docker, Azure)

- **Achieved 87% precision and reduced false positives by 32%** by optimizing weighted cross-entropy loss and fine-tuned SGD (momentum = 0.9, weight decay = 5e-4).
- **Deployed scalable, real-time inference** on Kubernetes with automated CI/CD pipelines.
- **Integrated structured and semi-structured patient data** from relational and NoSQL sources for comprehensive healthcare analytics.

Social Recommendation System -(Graph Neural Networks, PyTorch Geometric, Neo4j)

- **Improved recommendation precision by 22%** by implementing a GraphSAGE-based recommendation engine.
- **Reduced query execution time by 35%** through Neo4j integration and optimized graph storage.
- **Deployed scalable real-time inference** via FastAPI with graph-based data modeling for performance and scalability.

Education

M.S. Computer Science| Indiana State University, Terre Haute, IN, US

- Courses: Statistics, Artificial Intelligence, Matrices, Data Mining, Data Visualization (Tableau/Power BI)

B.S. Computer Science| Nirma University, Ahmedabad, GJ, India

- Courses: Deep Learning, Big Data Analytics (Hadoop), Applied Statistics, Calculus, Financial Management, Supply Chain

Future Goals

- Expand inference expertise into **State Space Models (SSMs)** to complement Transformer-based pipelines.
- Contribute to **hybrid model architectures** and **custom inference engines** for real-time multimodal intelligence.