SHREYAN SOOD
14215, Buffalo, New York, USA
shreyansood@gmail.com | +1 8583196576 | Portfolio | GitHub | LinkedIn

WORK EXPERIENCE

Machine Learning Engineer | Full Time **September 2024 – Present**
Prompt Inversion AI, Dover, Delaware

- Scaled an **LLM-driven** FastAPI backend with async agentic workflows, reducing response time by **60%** with **10x** concurrency.
- Implemented **RAG** based vector search algorithm with dynamic thresholding, enhancing document matching accuracy by **40%**.
- Orchestrated **Docker** containerization and **CI/CD** pipeline for zero-downtime deployment to **AWS EC2** with **95%** test coverage.
- Built the web application with Next.js/TypeScript and deployed to **AWS Amplify** with Auth0 integration for user authentication.

Analytical Scientist Intern | Internship **June 2023 – December 2023**
FICO, San Diego, California

- Built end-to-end monitoring system to detect model drift in a fraud-detection neural net using self-devised time-series algorithm.
- Validated the system for **15** key clients, triggering alerts for significant shifts in distribution and reduced false positives by **90%**.
- Developed an ETL pipeline to compute and visualize the distribution of terabyte-scale datasets, cutting processing time by **50%**.
- Ran statistical experiments to simulate and cluster drastic shifts in customer behavior patterns and sophisticated fraud schemes.

AI Engineer | Full Time **February 2022 – August 2022**
Collablens, Haryana, India

- Developed and deployed multiple **AI stations** for automated drop testing of flour packets. Integrated hardware with **cloud-based** modules for spillage detection, orientation checks and real-time analytics, leveraging GStreamer for **live video streams**.
- Secured a contract to deploy the system in **50 factories** and raised **$200,000** in funding, with investment from MIT Media Lab.
- Prototyped a versatile **Computer Vision System** for real-time defect detection in laser-engraved products, enabling automated quality control on the factory's assembly lines. Achieved **95%** accuracy and a mean inference time of **2.5 seconds per board**.

Machine Learning Intern | Internship **May 2021 – December 2021**
Hypertechpreneurs, Haryana, India

- Productionized a Vehicle Damage Detection Model utilizing Mask R-CNN for Instance Segmentation to automate inspections.
- Integrated model inference endpoints in web and mobile applications and developed a robust OCR system with **92%** accuracy.

EDUCATION

University of California San Diego **2022 – 2024**
Master of Science, Data Science *GPA (3.98/4.0)*
Teaching Assistant for DSC-261: Data Ethics, DSC-291: Statistical Models, DSE-250: Relational Data Models
Courses: *Causal Inference, Visual Learning, Search & Optimization, Scalable Systems, Data Management, Recommender Systems*

Delhi Technological University, New Delhi **2018 – 2022**
Bachelor of Technology, Mathematics and Computing Engineering *CGPA (8.73/10)*

PROJECTS

FRIES IN THE BAG AI **November 2024 – Present**

- Building full-stack AI-powered automated job application assistant using Django and React to streamline the application process.
- Implemented secure profile management system for storing and managing credentials, work history, and application preferences.
- Integrating RAG driven pipeline to generate tailored responses and cover letters based on job description and company research.

Rubik's Cube 3D Visualizer and Deep Reinforcement Learning Solver **July 2024 – September 2024**

- Developed NxN Rubik's Cube 3D visualizer and implemented Monte Carlo Tree Search algorithm optimized with a self-designed deep reinforcement learning ResNet. Achieved 71% solution rate for 9-move scrambles and sub-1 second solve times.

MediLoRA: LLM for medical Q&A with QLoRA **October 2023 – January 2024**

- Fine-tuned OpenHermes-2.5-Mistral-7B using Q-LoRA on 300M medical text tokens, improving PubMedQA and MedQA accuracy by over **20%**. Matched state-of-the-art 70B model performance on MMLU-Medical with just **0.05%** of the data size.

RESEARCH EXPERIENCE AND PUBLICATIONS

- Sood, S., Jain, T., Batra, N., Taneja, H.C. (2023). Black–Scholes Option Pricing Using Machine Learning. [ICDSA 2023]
- S. Sood and Y. Ahuja, "Selective Lossy Image Compression for Autonomous Systems." [STSIVA 2021]

TECHNICAL SKILLS

- **Programming :** Python, SQL, R, C++, MATLAB, JavaScript, HTML, CSS
- **Technologies :** Pandas, PyTorch, TensorFlow, Git, JAX, Hadoop, LangChain, AWS, PostgreSQL, PySpark, Docker, Next.js
- **Skills :** Data Science, Machine Learning, Data Engineering, Computer Vision, Language Processing, Data Analytics