SHREYAN SOOD
78704, Austin, TX, USA
shreyansood@gmail.com | +1 (858) 319-6576 | [Portfolio](#) | [GitHub](#) | [LinkedIn](#)

WORK EXPERIENCE

Machine Learning Engineer   September 2024 – Present
Prompt Inversion AI, Dover, Delaware

- Scaled an LLM-driven FastAPI backend with async design for job sourcing and email campaigns, cutting response time by **60%**.
- Developed an AI-powered resume matching algorithm using RAG based vector-search, improving selection accuracy by **40%**.
- Built and deployed Next.js web app on AWS Amplify for AI-driven job placement with authentication and real-time analytics.
- Set up AWS CI/CD pipeline (CodePipeline, ECR, EC2) for zero-downtime containerized deployments with **95%** test coverage.
- Built multi-agent spam classification workflows using six different AI agentic frameworks with human feedback retraining loop.
- Built unified Python framework for red teaming LLMs, implementing jailbreak algorithms to test AI safety and model alignment.

Analytical Scientist Intern   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 **75%**.
- Ran statistical experiments to simulate and cluster drastic shifts in customer behavior patterns and sophisticated fraud schemes.

AI Engineer   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   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

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 co-pilot using GPT 4o-mini to streamline the application process.

[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, JavaScript, C++, R, MATLAB, HTML, CSS
- **Technologies :** Pandas, NumPy, PyTorch, TensorFlow, Git, REST, AWS, PostgreSQL, Spark, Docker, Next.js, LangChain
- **Skills   :** Data Science, Machine Learning, Data Engineering, Computer Vision, Language Processing, Data Analytics