SHAHZAIB SAQIB WARRAICH

warraich@usc.edu | (213) 681-6140 | linkedin.com/in/shahzaib-saqib-warraich | https://shahzaib-s-warraich.github.io

EDUCATION

University of Southern California, Viterbi School of Engineering — **Los Angeles, CA**
M.S., Applied Data Science — *Jan 2024 - Present*

EXPERIENCE

Data Science Intern — **Burbank, CA**
Paramount Pictures — Jun 2025 - Present
- Leveraging petabytes of external and internal content engagement signals for content forecasting and production green lighting.
- Building GenAI powered explainability tools for business KPI analyses.

Graduate Research Assistant — **Los Angeles, CA**
Data, Interpretability, Language and Learning (DILL) lab, USC-NLP Group — Jan 2024 - Present
*A*dvisors: Gregory Yauney and Swabha Swayamdipta
- Currently interested in the intersection of LMs, Interpretability, Evaluation, Reasoning, and Alignment.
- Developed inference-time alignment frameworks to augment the factuality and reasoning ability of LMs, reducing hallucination across multiple benchmarks.

Research Engineer — **Redmond, WA**
Retrocausal — Dec 2022 - Dec 2023
- Engineered precision assembly solutions with up to **98%** detection accuracy for industry leaders such as Schneider Electric, Carrier, and Honda using cutting-edge object detection and activity recognition algorithms.
- Led the development of an automated object detection model training pipeline, which employed the cut, paste, and mix augmentation technique to generate synthetic dataset alongside labeled data, thereby reducing manual data labeling time by **5x**.

AI Engineer — **San Jose, CA**
Adlytic AI — June 2021 - Dec 2022
- Led R&D on model optimization and pruning, leveraging knowledge distillation and Nvidia TensorRT to reduce edge inference latency by **3x** with less than **1.5%** accuracy loss.
- Developed and deployed a cloud-based recommender system pipeline, achieving up to **97%** accuracy on unseen data.

INDUSTRY PROJECTS

Person Re-identification System | *PyTorch, Computer Vision, Nvidia TensorRT, FastAPI* — April 2022 - Dec 2022
- Leveraged facial recognition, super resolution, object detection, and multi-object tracking algorithms to design the model inference API. Utilized RTSP to perform object tracking on frames from live streams.
- Optimized object detection models to achieve an accuracy of up to **93%** with an IoU of nearly **0.77**.

Mobile App Recommender System | *PySpark, AWS, MongoDB, Redis, Docker* — June 2021 - Mar 2022
- Designed end-to-end system architecture on AWS (SageMaker, EMR, EC2, S3, SNS).
- Optimized matrix factorization models to generate collaborative and content based recommendations using Apache Spark.
- Stored and served recommendations using MongoDB and AWS managed Redis databases.
- Packaged and deployed application using Docker containers.
- Boosted the product's installation and client conversion rate by nearly **200%**.

RESEARCH PROJECTS AND PUBLICATIONS

- Sample, Align, Synthesize: Graph-Based Response Synthesis with ConGrs (NeurIPS 2025 Conference Submission)
- How Reliable is Language Model Micro-benchmarking? (NeurIPS 2025 Conference Submission)
- Sample, Align, Synthesize: Graph-Based Response Synthesis with ConGrs (COLM 2025 Workshop SCALR Submission)
- Residual balanced attention network for real-time traffic Scene Semantic Segmentation (IJECE 2023 Journal Submission)

LEADERSHIP

Viterbi Graduate Student Association (VGSA)

VP of Outreach Aug 2024 - Present

- Representing a student body of almost **2,000** students.
- Managing internal and external partnerships.
- Planning and organizing professional career development events.

SKILLS AND INTERESTS

Skills

Languages (Python, C/C++, SQL), Deep Learning Frameworks (PyTorch, Hugging Face, vLLM), Machine Learning Frameworks (Scikit-Learn, OpenCV), Data Analysis Frameworks (Pandas, Dask, PySpark), Data Visualization (Matplotlib, Seaborn, Plotly), Cloud (AWS Sagemaker, GCP Vertex AI), Others (Docker, Kubernetes, MongoDB, Git, LaTeX)

Interests

AI Ethics and Responsibility, Hackathons, User Experience (UX) Design