Mohammad Hasnain Raza

Los Angeles, CA
linkedin.com/in/hasnainraza03 | github.com/hasnainrazaa03

Email: razam@usc.edu
Mobile: (213) 994-5086
hasnainrazaa.vercel.app

Education

- **University of Southern California** — Los Angeles, CA
  Master of Science in Computer Science; ***GPA: 4.00*** — *Aug 2025 – Dec 2027*
  - Coursework: Algorithms, Database Systems, Programming System Design, Computer Networks, Machine Learning
- **RV College of Engineering** — Bengaluru, India
  Bachelor of Engineering in Aerospace Engineering; ***GPA: 3.86, Silver Medalist*** — *Aug 2018 – Jul 2022*
  - Coursework: Engineering Mathematics, C Programming, Computational Fluid Dynamics (CFD), Scientific Computing

Experience

- **Deloitte** — Bengaluru, India
  Technology Analyst — *Aug 2022 – Nov 2024*
  - Achieved 10x throughput increase by designing Pega PRPC SaaS with REST API orchestration for customer workflows across 35+ countries, processing 7,500+ creations and 12,500+ modifications.
  - Confirmed 44% latency reduction via Welch two-sample t-test ($n = 2,500$+ samples, $p < 0.05$), providing statistical evidence for workflow optimization effectiveness by accounting for unequal variances and non-normal execution time distributions.
  - Engineered LLM orchestration pipelines using Few-Shot and Chain-of-Thought prompting to automate complex workflow categorization, achieved 92% alignment with human-annotated labels via RLHF-driven feedback loops.
  - Standardized data governance protocols for domain-specific LLM training sets; implemented automated quality checks and PII-masking filters that reduced data preparation overhead by 30% while ensuring ethical AI compliance.
- **Defence Research and Development Organisation (DRDO)** — Bengaluru, India
  Research Intern — *Jan 2022 – Aug 2022*
  - Led a 4-member team to automate PyFluent-based transient CFD pipelines (k-ω SST, overset mesh), generating 1.2TB+ of high-fidelity aerodynamic datasets for training surrogate ML models and NLU-based physics emulators.
  - Conducted multi-variate regression analysis on CFD pressure fields to identify key aerodynamic predictors, engineered robust data-cleaning pipelines using Z-score outlier detection to ensure 99.8% data integrity for downstream AI training.
  - Developed automated MATLAB post-processing modules for terabyte-scale datasets, extracting high-dimensional feature vectors to validate flight-path trajectories against empirical sensor data with sub-millimeter precision.
- **Prana.ai** — Remote
  Founding Engineer — *Sep 2019 – Dec 2021*
  - Built end-to-end ML pipelines preprocessing and augmenting 5M+ MRI/CT volumes with normalization, patch extraction, and data balancing. Secured $50,000 pre-seed from First Round Capital.
  - Implemented depthwise separable CNN architecture for real-time 3D medical image segmentation, achieving $< 0.8s$ inference latency. Optimized inference pipeline enabling real-time diagnostic assistance.
  - Developed SO(3)-equivariant CNN-based super-resolution techniques using the e3nn library and rotation-equivariant convolutions to enhance image fidelity, achieving approximately 35% improvement in image clarity.

Key Projects

- **Project Vimaan – AI Voice Command NLU System** — Los Angeles, CA
  NLP/ML Pipeline with X-Plane Integration — *Sep 2025 – Present*
  - Architected an automated data-generation pipeline producing 30,000+ labeled examples optimized for joint intent-and-slot extraction, utilized schema-driven Python scripts and zero-shot AI-driven paraphrasing via Pegasus and FLAN-T5.
  - Fine-tuned DistilBERT (Transformer architecture) for multi-task joint intent-and-slot NLU, applied dynamic INT8 quantization to optimize the computational graph, achieving a 4x reduction in model size while preserving 98% of baseline accuracy.
  - Integrated the trained NLU model into the X-Plane simulator plugin architecture by developing custom interfaces with thread-safe inter-process communication (IPC) and text-to-speech (TTS) feedback.
- **USC Ledger – AI-Powered Financial Management Platform** — Los Angeles, CA
  LLM Integration & Data Engineering — *Aug 2025 – Present*
  - Architected a robust transactional data ingestion pipeline (React, Node.js, MongoDB) featuring a sequential reconciliation engine to guarantee 100% data consistency and high-fidelity structured inputs for downstream AI financial modeling.
  - Formulated high-throughput feature processing logic using fixed-precision arithmetic and 800ms debounced event throttling, optimizing concurrent data streams to strictly control API payload size and minimize latency during real-time LLM inference.
  - Deployed an LLM integration layer utilizing the Google Gemini API, designed scalable prompt engineering frameworks and context-window optimization techniques to automate financial anomaly detection and extract structured savings insights.

Skills

- **Languages**: Python (Advanced), SQL, R, MATLAB, Java, C/C++, JavaScript
- **AI/LLM Core**: Transformer Architectures, LLM Fine-tuning, Prompt Engineering, RLHF, NLP, BERT, T5
- **Frameworks & MLOps**: PyTorch, TensorFlow, Hugging Face, Scikit-learn, Docker, Model Quantization, Model Monitoring
- **Data Engineering**: Data Pipelines, ETL, Data Governance, Feature Engineering, Z-score Outlier Detection, Pandas, NumPy
- **Cloud & Tools**: AWS, Google Colab, Git, ServiceNow, Pega PRPC, Tableau