Shreyas Tulsi

US Citizen | shreyastulsi@ucla.edu | (925)-523-9257 | LinkedIn | Github

EDUCATION

University of California, Los Angeles (UCLA)

Bachelor of Science in Computer Science *Expected Graduation June 2026*

- **Related Coursework:** Object-oriented Programming (C++), Software Construction (Emacs, Bash, React), Operating Systems (C, GDB), Programming Languages (Lisp, Rust, OCaml, Java), Computer Networking, Web Applications

EXPERIENCE

Artificial Intelligence Researcher @ UCLA Data Mining Lab

Python · PyTorch · Hugging Face · Transformers · CUDA *July 2025 – Present*

- Lab Focus: Accelerating LLM inference through speculative draft-then-verify decoding methods
- Co-developed a speculative decoding framework, titled *Octopus*, with router-based expert head selection to generate multi-token drafts in parallel, integrating tree attention and achieving 3–4× speedups
- Designing and running controlled studies on diverse training strategies, evaluating loss functions, pre-head inclusion, and positional encoders to maximize tokens-per-second throughput and draft-based model alignment

Software Development Engineer Intern @ Amazon

Java · Python · Android(Kotlin) · DynamoDB · SQL · XGBoost · AWS Sagemaker · Amazon Bedrock *June 2025 – Sept 2025*

- Deployed an AWS SageMaker XGBoost model on historical driver data with temporal and severity features to flag risky driving behaviors, boosting precision by 19% and reducing alert noise by 28%
- Integrated the ML-enhanced driver violation detection model into the existing Java backend, designing a new event flow with AWS CDK, Lambda, and API Gateway, and implementing CloudWatch logging for observability
- Introduced personalized LLM-based driver violation summaries via Amazon Bedrock into existing Android mobile app

Software Engineer Intern @ Anvi Cybernetics

RAG · LangChain · AWS · Flask *September 2024 – May 2025*

- Led 8 interns in developing a RAG-powered application that converts 1–2 page synopses into full clinical trial protocols
- Applied fine-tuned Hugging Face models, LangChain pipelines, and vector database technologies to achieve 70–80% accuracy across protocol sections, enabling clinicians to save nearly three weeks in document creation
- Developed a Flask-based web platform on AWS that enabled clinicians to upload PDF synopses, integrated with S3 storage and the hosted RAG pipeline to automatically generate complete and professional-grade protocol documents

Machine Learning Researcher @ UCLA Goodman-Luskin Microbiome Lab

R · MaAsLin2 · QIIME2 · MATLAB *October 2023 – June 2024*

- Applied CNNs with MATLAB CONN toolbox and Structural Equation Modeling (SEM) to analyze fMRI brain scans for cognitive–behavior insights, achieving 85% classification accuracy and explaining 40% variance in SEM models
- Built 16S pipeline: trimmed FASTQ files, computed alpha diversity (QIIME), ran PCoA for beta diversity, and performed bacterial abundance modeling (MaAsLin via R), identifying 150 taxa with significant association
- Applied ANOVA and logistic regression to evaluate how perceived discrimination drives substance cravings, uncovering 12 significant predictors and achieving 76% model accuracy with a pseudo-R^2 of 0.21

COMPETITIONS

- Cybersecurity: top-1% in picoCTF (164/18 k, international individual) and top-2% in CyberPatriot (31/2 k teams)
- United States of America Computing Olympiad (USACO) – Gold Division Qualification

PROJECTS

Production-Deployed AI Nutrition & Meal Platform *March 2025 – June 2025*

- Launched a cloud-hosted Progressive Web App for pantry inventory, macro tracking, and AI recipe suggestions, enabling users to scan items via a WASM barcode reader with full offline support
- Engineered the React/Vite frontend and Node/Express + PostgreSQL backend with secure JWT auth, Redis caching, Workbox/IndexedDB, and Docker-based CI/CD deployed to Google Cloud Run

Attention-Informed LLM Pipeline for Image Ad Optimization *January 2025 – April 2025*

- Developed an LLM-powered Photoshop API pipeline that applies demographic-tuned color palettes, auto-recolors images, and repositions assets using attention maps, then A/B ranks variants to improve targeting accuracy by 30%

RAG-Enabled LLM Platform for Multimedia Learning *May 2024 – September 2024*

- Built a personalized edtech platform where users upload PDFs or YouTube links, with the system processing content through a LangChain RAG pipeline with a fine-tuned Hugging Face model to generate relevant & adaptive questions