Shriya Sravani Yellapragada

+1 (408) 413-6948 | shriyasravani07@gmail.com | LinkedIn | Github | Portfolio

PROFESSIONAL EXPERIENCE

Machine Learning Engineer, **Stealth Startup** | **Santa Clara, CA** **Oct 2025 - Present**

- Led the development of a multilingual video delivery pipeline on Google Cloud Platform (GCP), integrating Mux Video API and Google Veo 3 for AI-generated video content, deploying **30+** educational videos across banking, insurance, childcare, and healthcare domains in 4 languages (English, Spanish, Portuguese, French), expanding application reach by **4x** across blue-collar workforce segments.
- Engineered an automated FastAPI-based upload workflow that eliminated manual video ingestion, dynamically routing language-specific content and metadata to expand platform accessibility across diverse user demographics.
- Optimized video content generation by implementing a Gemini Flash 2.5-powered prompt enhancement system, producing high-quality scripts for NotebookLM video synthesis and streamlining content creation workflow.
- Enhanced code quality and maintainability using Augment AI for automated refactoring, streamlining the video processing pipeline and improving deployment efficiency.

Student AI Researcher, **UCSC, Adobe** | **Santa Clara, CA** **May 2025 - Dec 2025**

- Engineered experiments evaluating LLM adherence to SQL reserved-keyword rules, generating **500+** structured samples that strengthened dataset reliability and reduced manual annotation time by nearly 18 researcher-hours weekly.
- Built an automated compliance-verification pipeline using 520 BIRD queries, improving prompt consistency across multilingual inputs and accelerating evaluation throughput from 40 to 210 processed instructions/hour.
- Developed a multilingual NLLB-200 workflow translating **500+ SQL prompts**, reducing cross-lingual ambiguity and cutting overall translation latency from 1.8 seconds to 320 milliseconds.
- Implemented a multi-model NER ensemble detecting **364+ entities** and caching Wikipedia lookups, decreasing network-fetch overhead by 80% and avoiding approximately $140 monthly API overage leakage.

Researcher And Member, **University Of California - Santa Cruz** | **Santa Cruz, CA** **Mar 2025 - Oct 2025**

- Designed a custom autonomous-vehicle simulation environment by extending CARLA with OpenAI Gym, enabling scalable RL experimentation and reducing environment reset time from 12 seconds to 1.9 seconds.
- Engineered motion-planning and sensor-fusion models that improved obstruction-avoidance accuracy to **90%**, eliminating repeated collision loops and saving nearly 40 GPU-hours across training experiments.
- Benchmarked PPO, A2C, and DQN policies across **50+** simulated episodes, improving policy stability and reducing convergence requirements.
- Presented findings as part of AV research to **120+** symposium attendees, providing reproducible RL benchmarks that supported cross-lab experimentation and influenced two subsequent AV simulation studies within the department.

Student Researcher, **University Of California - Santa Cruz** | **Santa Clara, CA** **Oct 2024 - Dec 2024**

- Built a multilingual fact-check retrieval system analyzing **5k+** social-media posts, improving top-k relevance and eliminating 3k noisy tokens using language-aware preprocessing and similarity-based filtering.
- Optimized Logistic Regression, BiLSTM-Attention, and XLM-RoBERTa models to raise cross-lingual retrieval accuracy to 0.87 while moderating inference latency.
- Hyperparameter-tuned XLM-RoBERTa to reach **0.78** standalone accuracy, strengthening multilingual generalization and stabilizing performance for low-resource language groups across three geographic regions.
- Collaborated with cross-disciplinary researchers to automate preprocessing workflows, reducing dataset cleaning time from a full workday to 35 minutes for each 1k-sample multilingual batch.

EDUCATION

University Of California - Santa Cruz (UCSC) - *Master's*, *Natural Language Processing* **Dec 2025**

Keshav Memorial Institute Of Technology (KMIT) - *Bachelor's*, *Computer Science & Engineering (AI & ML)* **Jun 2024**

PROJECTS

LLM-as-a-Judge

- Developed an all-to-all LLM evaluation framework using 10 models judging each other on ELI5 and WritingPrompts tasks, identifying biases like verbosity preference through structured reasoning scores.

Multilingual Fact-Checked Claim Retrieval System

- Built a multilingual fact-check retrieval system improving cross-lingual accuracy to 0.87 and refining XLM-RoBERTa to 0.78 standalone accuracy while automating preprocessing across diverse language datasets.

Slot Tagging On Movie Dataset

- Trained and benchmarked MLP, LSTM, and CNN architectures with diverse embeddings, achieving 89% slot-tagging accuracy using a BiLSTM model optimized for movie-domain natural-language queries.

SKILLS

Programming : Python, Java, C/C++, SQL, Bash, HTML, CSS

Machine Learning : Scikit-learn, XGBoost, Logistic Regression, Clustering, Feature Engineering

Deep Learning : PyTorch, TensorFlow, BiLSTM, CNN, Transformers, Attention Models

NLP & LLMs : XLM-RoBERTa, BERT, NLLB-200, spaCy, Flair, NER, Language Modeling, MT, RAG, Prompt Engineering

AI Evaluation : LLM-as-a-Judge frameworks, Structured Reasoning, Model Benchmarking, Prompt-based Evaluation

Reinforcement Learning : PPO, A2C, DQN, OpenAI Gym, CARLA Simulator, Policy Optimization

Data Processing : NumPy, Pandas, Regex Pipelines, JSON Caching, Data Cleaning Automation

Cloud & DevOps : AWS, Azure, Kubernetes, Git, GitHub Actions

Databases : MySQL, SQLite, Vector Stores (FAISS / approximate search)

Tools : Streamlit, Jupyter, CARLA, Docker, FastAPI, REST APIs, Java Servlets

CERTIFICATIONS

Foundations: Data, Data, Everywhere, Google

Advanced RAG With Vector Databases, IBM

GenAI Job Simulation, BCG

Software Engineering Job Simulation, Goldman Sachs