Shriya Sravani Yellapragada

Machine Learning Engineer

+1 (408) 413-6948 | shriyasravani07@gmail.com | linkedin | github | portfolio

PROFESSIONAL EXPERIENCE

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

- Engineered experiments to evaluate LLM adherence to SQL reserved-keyword rules, generating 500+ structured samples that enhanced dataset reliability and reduced manual annotation time by nearly 18 researcher-hours weekly.
- Built an automated compliance-verification pipeline using 520 BIRD queries, ensuring 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, minimizing 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, reducing network-fetch overhead by 80% and preventing approximately $140 monthly API overage.

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

- Engineered a custom autonomous-vehicle simulation environment with Python, extending CARLA using OpenAI Gym for scalable RL experimentation, reducing environment reset time from 12 seconds to 1.9 seconds.
- Developed motion-planning and sensor-fusion models, enhancing obstruction-avoidance accuracy to 90%, thereby eliminating repeated collision loops and saving nearly 40 GPU-hours across training experiments.
- Conducted benchmarking of PPO, A2C, and DQN policies across 600+ simulated episodes, improving policy stability and reducing convergence requirements from 180k to nearly 50k training steps.
- Presented AV safety research to 120+ symposium attendees, sharing reproducible RL benchmarks that facilitated 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**

- Developed a multilingual fact-check retrieval system, enhancing customer support efficiency by analyzing 40k+ social-media posts, improving top-k relevance, and eliminating 31k 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 reducing inference latency from 900ms to 210ms across six languages, contributing to improved user experience.
- Hyperparameter-tuned XLM-RoBERTa to achieve 0.78 standalone accuracy, enhancing multilingual generalization and stabilizing performance for low-resource language groups across three geographic regions, ensuring reliable service delivery.
- Collaborated with cross-disciplinary researchers to automate preprocessing workflows, cutting dataset cleaning time from a full workday to 35 minutes for each 10k-sample multilingual batch, streamlining operations.

EDUCATION

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

Keshav Memorial Institute Of Technology - *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

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

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, Intercom, GitHub, Jira

CERTIFICATIONS

Foundations: Data, Data, Everywhere, Google

Advanced RAG With Vector Databases, IBM

Transformer-Based NLP, Nvidia

GenAI Job Simulation, BCG

Software Engineering Job Simulation, Goldman Sachs