Shreyas Rajesh

shreyasrajesh38@g.ucla.edu | +(424)-440-9948 | Linkedin | github | homepage | scholar

Education

University of California, Los Angeles(UCLA) 2023 - Present
Ph.D. in Electrical and Computer Engineering Advisor: Prof. Vwani Roychowdhury

Research Focus: Language Modelling, Information Retrieval and Memory Systems for LLMs

University of California, Los Angeles(UCLA) 2021-2023
M.S. in Electrical and Computer Engineering

Selected Papers

- **Shreyas Rajesh**, Pavan Holur, Chenda Duan, David Chong, Vwani Roychowdhury. "Beyond Fact Retrieval: Episodic Memory for RAG with Generative Semantic Workspaces".
  AAAI 2026 (Oral)
  NeurIPS 2025 *Workshop on Language, Agents and World Models* (Spotlight).
- Pavan Holur*, Kenneth C. Enevoldsen*, **Shreyas Rajesh**, Lajoyce Mboning, Thalia Georgiou, Louis-S. Bouchard, Matteo Pellegrini, Vwani Roychowdhury. "Embed-Search-Align: DNA Sequence Alignment using Transformer models".
  Bioinformatics 2025.
- Zhe Fei*, Mehmet Yigit Turali*, **Shreyas Rajesh***, Xinyang Dai, Huyen Pham, Pavan Holur, Yuhui Zhu, Larissa Mooney, Yih-Ing Hser, Vwani Roychowdhury. "Customizing Open Source LLMs for Quantitative Medication Attribute Extraction across Heterogeneous EHR Systems".
  NeurIPS 2025 *Workshop on GenAI for Health.*
- Pavan Holur, **Shreyas Rajesh**, David Chong, Vwani Roychowdhury. "Creating an AI Observer: Generative Semantic Workspaces".
  arXiv (2024).

Work Experience

Nvidia June 2025 - Dec 2026
PhD Intern - Large Language Models

- Focusing on improving on-device performance for small language models (SLMs) through retrieval (RAG) and other techniques.
- Building agentic systems with small language models (SLMs) that can operate efficiently on edge devices while maintaining high performance.

Roychowdhury Group, UCLA Feb. 2023 - Current
Graduate Student Researcher

- Ph.D. Researcher with Prof. Vwani Roychowdhury on problems in NLP and Brain-inspired AI.
- Currently working with and fine-tuning Large Language Models (LLMs) like LLaMA and Mistral 7B using parameter efficient methods like LoRA for NLP tasks mainly focused on situation modelling and building workspaces for LLMs.
- Also trying to adapt advances in language modeling to build representation models for other types of data like genomic sequences and brain signals by using encoder style transformer models.

Nvidia May 2024 - Sept 2024
PhD Intern - Large Language Models

- Working on finetuning and deploying large language models like LLaMA and Phi.
- Further my goal is to reduce cost of performing particular tasks by over 70% through finetuning while maintaining the same level of performance.
- Also working on deploying these models on edge devices and optimizing them for performance and memory constraints.

Elseware Feb 2023 - June 2023
Machine Learning Engineering Intern

- Adding a full feature on Elseware's MSTAR online platform to summarize and classify articles.
- This involves collecting and cleaning information from thousands of financial articles and documents to create a well curated dataset.
- Using the developed dataset to finetune large language models like GPT-3 to perform classification and summarization tasks.
- Also created a pipeline to perform the same tasks using the ChatGPT API from OpenAI using additional tools like guardrails and langchain.

Airprobe Oct. 2020 - Mar. 2021
Deep Learning and Computer Vision Engineer

- Digitized entire solar plants as part of the Computer Vision team.
- Trained State of the art Deep Learning Models YOLOv4 and Detectron to detect barcodes from drone images of these plants.
- Achieved a 50% reduction in both time and manpower required to digitize entire plants.

Technical Skills

Programming and Scripting Languages : Python, C++, MATLAB, Bash

Frameworks and Packages: PyTorch, Huggingface, Pandas, Scikit-learn, Scikit-image, OpenCV, NumPy

Operating Systems : Linux, MacOS, Windows