Sreeram Vennam

https://vnnm404.github.io/
vennam404@gmail.com
https://github.com/vnnm404/

Education

2021–2025 **B.Tech. with Honors in Computer Science**
International Institute of Information Technology, Hyderabad GPA: **9.54**/10
RELEVANT COURSEWORK: Topics in Deep Learning (Graph Neural Networks), Advanced NLP, Deep Learning, Statistical Methods in AI, Computer Vision, Distributed Systems, Operating Systems and Networks, Linear Algebra, and Probability and Statistics.

Work Experience

Monsoon 2024 **Teaching Assistant**, *Statistical Methods in AI* by Dr. Ravi Kiran
Evaluated assignments and exams for over 150 students in topics including linear algebra, probability and statistics, backpropagation, CNNs, and RNNs and took part in preparing assignments and seminal assessments.

Summer 2024 **Software Development Intern**, *Google*
Produced a connector to transfer large scale data from SQLServer into BigQuery for data migration. Introduced a novel data partitioning mechanism which led to faster parallel transfers. Received a PPO for my outstanding performance.

Spring 2024 **Research Intern**, *Subtl.ai*
Constructed a pipeline for RAG over audio logs for automated form filling. Incorporated SOTA audio transcription models and improved pipeline performance by 44% through chain-of-thought prompting.

Research Experience

2024 **Reasoning in LLMs**, *Czech Technical University in Prague*
MENTOR: David Herel
Worked on investigating the properties of Thinking Tokens and Pause Tokens and explained their underperformance through gradient analysis [1]. Currently working on a novel method to compress chain-of-thought tokens to reduce the memory footprint of large reasoning models.

2024 **LLM Vocabulary Compression for Low-Compute Environments**, *Precog*
MENTOR: Dr. Ponnurangam Kumaraguru
Introduced a novel approach to reduce the memory footprint of the logits tensor in LLMs to $\sqrt{V}$ theoretical complexity enabling the pre-training of LLMs in low-compute environments. [2]

2024 **Interpreting OCR in Visual Language Models**, *Precog*
MENTOR: Dr. Ponnurangam Kumaraguru
Fascinated by how humans process text visually, I sought to explore the existence of

textual information within the image encoder of CLIP. Found strong evidence for textual semantics hidden within the image pooling embedding for rendered text that is robust to fonts and keyword matching. [3]

2024 **Automated Data Exploration in Agentic Systems**, *Google 20p Project*
MENTOR: Sai Charan Tej
Silent model collapse in automated data exploration led to stale analysis regardless of the data diversity. I improved generated analysis through prompting the system to hypothesize the data origin, and then use domain specific analysis tools resulting in rich analysis.

2024 **Integrating Algebraic Topology into Neural Networks**, *Precog*
MENTOR: Dr. Charu Sharma
Produced a novel framework to incorporate higher order interactions in GNN explainers through cell complexes – significantly improving performance on both real world and synthetic datasets. [4]

2023 **Moral Inconsistency in LLMs**, *University of Maryland, Baltimore County*
MENTOR: Dr. Manas Gaur
First to show that LLMs were inconsistent in providing moral advice despite semantically identical prompts. Introduced a simple prompting strategy using Rule of Thumbs resulting in more consistent LLMs. [5]

Selected Publications

[1] **Rethinking Thinking Tokens: Understanding Why They Underperform in Practice**
Sreeram Vennam, David Valente, David Herel, Ponnurangam Kumaraguru
Under Review

[2] **Emergence of Text Semantics in CLIP Image Encoders**
Sreeram Vennam*, Shashwat Singh*, Anirudh Govil, Ponnurangam Kumaraguru
UniReps Workshop @ NeurIPS 2024

[3] **LLM Vocabulary Compression for Low-Compute Environments**
Sreeram Vennam, Anish Joishy, Ponnurangam Kumaraguru
Machine Learning and Compression Workshop @ NeurIPS 2024

[4] **Higher Order Structures For Graph Explanations**
Akshit Sinha*, **Sreeram Vennam***, Charu Sharma, Ponnurangam Kumaraguru
The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

[5] **SaGE: Evaluating Moral Consistency in Large Language Models**
Vamshi Krishna Bonagiri, **Sreeram Vennam**, Priyanshul Govil, Ponnurangam Kumaraguru, Manas Gaur
***Oral (Top 15%)** at The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*

Selected Projects

NanoTube, https://github.com/vnnm404/nanotube
YouTube from scratch in a distributed setting which implements the basics of distributed video streaming such as chunking and non-byzantine fault tolerance.

Gradf, https://github.com/vnnm404/gradf
Implements reverse mode automatic differentiation in C.

PyTorch Transformer, https://github.com/vnnm404/pytorch-transformer
A clean and shape annotated implementation of "Attention is All You Need" in pure PyTorch.

Konachan Image Scraper, https://github.com/vnnm404/konachan_dl
An asynchronous image scraper for the website Konachan resulting in a performance gain of **300%** over synchronous scraping.

Honours and Achievements

2024 SAGE [4] Accepted for **Oral** Presentation (Top 15%) at LREC-COLING 2024.

2024 Research Award at IIIT, Hyderabad (Only awarded to 7 CSE undergraduates)

2021–2024 3x Dean's List 1 (Top 5%), 1x Dean's List 2 (Top 10%), 1x Dean's List 3 (Top 15%)

2021 Rank **1001** in JEE Advanced out of 151,192 candidates.
Rank **367** in JEE Mains out of 1.12 million applicants.

Academic Service and Outreach

REVIEWER: Neural Compression @ Neurips 2024, Emergency Reviewer for CIKM 2024

UNIVERSITY GROUPS: Open Source Developers Group (Technical Member), Adventure Club (Founder)

VOLUNTEER WORK: Produced content for an LLM workshop for students of BVRIT, Hyderabad and I-HUB, IIIT Hyderabad. Produced content for the Responsible AI ACM India Summer School. Worked on the website for a local climbing gym.

TALKS: Presented SAGE at RnD Showcase, IIIT Hyderabad 2024. Presented FORGE at NPTEL F24: Responsible & Safe AI.