Ravisri (Ravi) Valluri

+1 848-702-6439 | ravisrivk@gmail.com | linkedin.com/in/ravi99 | ravi-vk.github.io | Los Angeles, CA

Education

University of California, Los Angeles (UCLA) **Sep 2025 – Present**
Ph.D in Computer Science (Advisor: Aditya Grover)
Distinctions: 2025 Graduate Dean's Scholars Award ($14,500) for highly recruited admitted students

Indian Institute of Technology, Madras (IIT Madras) **July 2019 – July 2023**
B.Tech in Computer Science & Engineering (Advisor: C. Chandra Sekhar) **GPA: 9.51 / 10.00**
Distinctions: 2021 S. Subramanian Prize for the highest first-year GPA (9.98/10.00)

Publications

- **Ravisri Valluri**, Akash Kumar, Kushal Dave, Amit Singh, Jian Jiao, Manik Varma, Gaurav Sinha.
  Scaling the Vocabulary of Non-autoregressive Models for Fast Generative Retrieval
  Proceedings of KDD '25, Research Track. DOI 10.1145/3690624.3709330

Relevant Experience

Research Fellow — Microsoft Research, India (with Profs. Gaurav Sinha & Manik Varma) **July 2023 – June 2025**

- Developed low-latency, accurate generative retrieval models by scaling vocabularies and optimizing inference algorithms; deployed with Microsoft Advertising across 150+ markets, generating tens of millions in revenue.
- Initiated and led the Extreme Vocabulary Project, which grew into a large research initiative recognized by leadership, including Microsoft AI CEO Mustafa Suleyman, and presented at the MSR India Academic Summit.
- Introduced a novel non-autoregressive (NAR) generative document retrieval model achieving 70x faster GPU inference while maintaining or improving accuracy over autoregressive (AR) baselines.
- Scaled vocabularies to 5M tokens (100x larger than BERT's 50K) with multi-word entities and phrases, improving retrieval accuracy and representation efficiency, and optimized inference via self-normalizing SoftMax and a learned shortlisting method for beam search efficiency.
- Built high-throughput training and inference pipelines for multi-billion parameter generative retrieval models, improving efficiency and scalability up to trillion-token datasets.
- Published at KDD 2025 and presented at the ICML 2024 SPIGM workshop; patents pending.

Undergraduate Researcher — IIT Madras (with Prof. C. Chandra Sekhar) **Jan 2023 – May 2023**

- Developed a multimodal architecture that projects images into GPT-2's embedding space for open-ended Visual Question Answering, enabling answer and explanation generation from text-only pretrained models.
- Improved visual reasoning by grounding image and text representations with ConceptNet knowledge graphs and a three-stage pretraining pipeline leveraging image-captioning data to transfer learning to VQA.

Software Engineering Intern — Microsoft Development Center, India **May 2022 – July 2022**

- Developed an ML inference system for seamless task/model switching, building and deploying a REST API using Python, Flask, SQL, and Azure Machine Learning.

Technical Skills

- Experience with information retrieval, large language models, extreme classification, and efficient inference, using Python, C++, SQL, PyTorch, ONNX, and DeepSpeed.