Vinamra Benara

Final-Year CS PhD Student
Sky Lab (formerly RISE Lab)
University of California, Berkeley

https://people.eecs.berkeley.edu/ vbenara/
Office Address: *465 Soda Hall, Berkeley, CA 94720*
vbenara@cs.berkeley.edu

Education

University of California, Berkeley *2019 - Present*
Doctor of Philosophy in Computer Science

- **Adviser: Prof. Ion Stoica**
- GPA 3.8/4
- **Interests: LLMs, AI, Systems**

International Institute of Information Technology, Hyderabad (IIIT-H) *2013 - 2018*
Bachelors (Hons.) and Masters (by Research) in ECE

- Advisers: Suresh Purini, Uday Bondhugula
- GPA 8.81/10, Department Rank 2

Publications

Crafting Interpretable Embeddings by Asking LLMs Questions (NeurIPS'24) *2024*
Vinamra Benara, Chandan Singh, John X. Morris, Richard Antonello, Ion Stoica, Alexander G. Huth, Jianfeng Gao.
Proceedings of the 38th Annual Conference on Neural Information Processing Systems **NeurIPS'24 (link)**.
OSS available ***here****.*

RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture (Microsoft Research) *2024*
Angels Balaguer, **Vinamra Benara**,...,Swati Sharma, Vijay Aski, Ranveer Chandra.
Preprint available ***here****.*

NumS: Scalable Array Programming for the Cloud *2022*
Melih Elibol, **Vinamra Benara**, Samyu Yagati, Lianmin Zheng, Alvin Cheung, Michael I. Jordan, Ion Stoica.
OSS available ***here****, Preprint available* ***here****.*

Bitwidth customization in image processing pipelines using interval analysis and SMT solvers *2020*
Suresh Purini, **Vinamra Benara**, Ziaul Choudhury, Uday Bondhugula.
Proceedings of the 29th International Conference on Compiler Construction ***CC'20*** **(link)**

Synthesizing power and area efficient image processing pipelines on FPGAs *2018*
Vinamra Benara, Ziaul Choudhury, Suresh Purini, Uday Bondhugula.
Preprint available ***here****.*

Accurus: A Fast Convergence Technique for Accuracy Configurable Approximate Adder Circuits *2016*
Vinamra Benara, Suresh Purini.
Proceedings of IEEE Computer Society Annual Symposium on VLSI ***(ISVLSI'16)****, pp. 577-582.* **(link)**

Research Experience

Microsoft Research, Student Researcher

- **Interpretable Embeddings** *Jan '24- ongoing*
  - Co-led the project on making LLM embeddings interpretable and enabling their application in critical domains such as neuroscience.
  - Resulted in a NeurIPS publication.
- **Copilots and Agents for M365 platform**
  - Designing agentic frameworks for various applications on the M365 platform.

Microsoft Research, Research Internship

- **Domain Adaptation for LLMs** *May '23- Aug '23*
  - Led the project from scratch. My work was the first to demonstrate the effectiveness of fine-tuning LLMs for knowledge injection.
  - It got widely covered on twitter etc with more than 500k impressions. -
  - Paper: RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture

Amazon CoreAI, Visiting Researcher

- **Distributed Probabilistic Inference on Ray** *Aug '20- Jan '21*
  - My work involved designing distributed probabilistic learning algorithms by integrating Ray with Amazon's internal tool called Clay.

RISE Lab/Sky Lab, PhD Student

- **NumS: Scalable Array Programming for the Cloud (OSS)** *Feb '21- Apr '22*
  - Co-led the design and development of NumS, which is a library that translates Python and NumPy to optimized distributed systems code. Project supervised by Ion Stoica and Michael Jordan.
  - Core maintainer on Github **(here)**
- **Fault Tolerant Distributed Data Parallel Training on Ray** *Oct '20- May '20*
  - Worked on designing a distributed data parallel training library that can run efficiently on unreliable instances like spot instances without severe overhead in cases of node failure, and automatic failure mitigation.
  - Reduction of overhead up to 10x.

Carnegie Mellon University (Pittsburgh), Research Assistant

- **Ultra low latency AR/VR headset prototype** *May '18- May '19*
  - Reduced motion-to-photon latency below **8 ms @ 240 frames per second.**
  - Designed an end-to-end display pipeline on an FPGA and reduced the latency of various vision algorithms.
- **Programmable Automotive Headlights**
  - Detects rain drops using a 1000 FPS camera and blocks light falling on the rain drops for improved visibility.
  - Designed a low latency communication infrastructure for inter headlight communication for various display pipelines.

Scholarships & Awards

- UC Berkeley CS Department Fellowship 2019.
- Received admits from Stanford CS, Berkeley CS and CMU ECE for PhD admission cycle - 2019.
- Finalist for Qualcomm Innovation Fellowship India, 2018 (Top 20 teams from India).
- Dean's Research list for excellence in undergraduate research for academic year 2015-2016.
- Dean's list I for Excellence in Academics (Top 5%) in 2^{nd}, 7^{th} and 8^{th} semester at IIIT-Hyderabad.
- Secured 99.68 percentile in All India Engineering Entrance Examination 2013 among 1.1M candidates.