

# Sriram Balasubramanian

📍 College Park, MD   ✉ sriramb@umd.edu   🌐 www.sriram.live   in sriramb1998   📄 Google Scholar

## Education

<b>University of Maryland, College Park</b> <i>PhD in Computer Science (advised by Prof. Soheil Feizi <a href="#">🔗</a>)</i>	<i>August 2023 – December 2026 (Expected)</i> GPA: 4.0
<b>University of Maryland, College Park</b> <i>MS in Computer Science (advised by Prof. Soheil Feizi <a href="#">🔗</a>)</i>	<i>August 2021 – May 2023</i> GPA: 4.0
<b>Indian Institute of Technology, Bombay</b> <i>B. Tech (Hons) in Computer Science &amp; Engineering (advised by Prof. Sunita Sarawagi <a href="#">🔗</a>)</i>	<i>August 2016 – May 2020</i> GPA: 9.56/10.0

## Work Experience

<b>Research Fellow (Incoming)</b> <i>FAR.AI</i>	<i>Berkeley, CA May 2026 – August 2026</i>
<b>Research Intern</b> <i>Adobe (Document Intelligence Lab)</i>	<i>San Jose, CA May 2025 – August 2025</i>
<ul style="list-style-type: none"> <li>Created a recipe to curate data and <i>post-train LLMs</i> using SFT and RL based algorithms to generate <i>citations</i> to answers by utilizing a decomposition based strategy. Preprint available <a href="#">here <a href="#">🔗</a></a>.</li> <li>Designed a new prompting strategy and a prompt optimization pipeline to improve attribution methods in Acrobat AI assistant by up to 10%.</li> </ul>	
<b>AI Researcher</b> <i>RelAI <a href="#">🔗</a></i>	<i>College Park, MD June 2023 - August 2023 June 2024 - August 2024</i>
<ul style="list-style-type: none"> <li>Developed and designed multiple applications to enhance vision model reliability for RelAI</li> <li>Involved in the development of RelAI, contributing from the inception stage through planning, execution, and deployment phases.</li> </ul>	
<b>Research Intern</b> <i>Comcast</i>	<i>Washington D.C. June 2022 - August 2022</i>
<ul style="list-style-type: none"> <li>Investigated the effectiveness of transfer learning in deep neural networks in the low resource regime (when the target domain has very limited data).</li> <li>Devised non-neural methods which could outperform both traditional collaborative filtering methods and neural networks in this regime.</li> </ul>	
<b>Research Fellow</b> <i>Microsoft Research</i>	<i>Bangalore, India August 2020 – August 2021</i>
<ul style="list-style-type: none"> <li><b>Predicting e-mail arrivals and reads:</b> Built machine learning models to predict e-mail arrivals and reads from user type and history of arrivals/reads to improve cache hit rates.</li> <li><b>Simulating network paths using ML:</b> Built machine learning models to simulate internet paths using static network traces</li> </ul>	
<b>Intern</b> <i>Tower Research Capital</i>	<i>Gurgaon, India May 2019 – August 2019</i>
<ul style="list-style-type: none"> <li>Developed tools to analyze trading and market logs, and compute and display metrics for various trading strategies and securities.</li> </ul>	

## Publications and Preprints

**Retrieval is Enough: Training-Free Interpretability with a Tool-Using Agent**

S Balasubramanian, S Feizi

*Under review at NeurIPS, ([link here](#)) [🔗](#), 2026*

**Circuit Oracle: Automating Attribution Graph Analysis via Natural-Language Queries**

HK Tan\*, S Kabir\*, S Agrawal, S Chereddy, S Balasubramanian

*Mentoring at SPAR, In Progress ([link here](#))* [🔗](#), 2026

### **Decomposition-Enhanced Training for Post-Hoc Attributions In Language Models**

**S Balasubramanian**, S Basu, K Goswami, R Rossi, V Manjunatha, R Santhosh, R Zhang, S Feizi, N Lipka  
*European Chapter of the Association for Computational Linguistics, Main*, 2026

### **A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models**

**S Balasubramanian**, S Basu, S Feizi

*Empirical Methods in Natural Language Processing (EMNLP), Findings*, 2025

### **Decomposing and Interpreting Image Representations via Text in ViTs Beyond CLIP**

**S Balasubramanian**, S Basu, S Feizi

**Spotlight** at *Mechanistic Interpretability Workshop, ICML*, 2024

*Advances in Neural Information Processing Systems (NeurIPS)*, 2024

### **Exploring Geometry of Blind Spots in Vision Models**

**S Balasubramanian\***, G Sriramanan\*, VS Sadasivan, S Feizi

**Spotlight** at *Advances in Neural Information Processing Systems (NeurIPS)*, 2023

### **Towards Improved Input Masking for Convolutional Neural Networks**

**S Balasubramanian**, S Feizi

*IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023

### **What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?**

**S Balasubramanian\***, N Jain\*, G Jindal\*, A Awasthi, S Sarawagi

*Rep4NLP Workshop @ Annual Meeting of the Association of Computational Linguistics (ACL)*, 2020

### **Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks**

VS Sadasivan, A Kumar, **S Balasubramanian**, W Wang, S Feizi

*Transactions on Machine Learning Research (TMLR)*, 2025

Media coverage at [Washington Post](#) [🔗](#) , [Wired](#) [🔗](#) , [TechSpot](#) [🔗](#) , [New Scientist](#) [🔗](#)

### **Rethinking Copyright Infringements in the Era of Text-to-Image Generative Models**

M Moayeri, **S Balasubramanian**, S Basu, P Kattakinda, A Chegini, R Brauneis, S Feizi

*International Conference on Learning Representations (ICLR)*, 2025

### **Gaming Tool Preferences in Agentic LLMs**

K Faghil, W Wang, Y Cheng, S Bharti, G Sriramanan, **S Balasubramanian**, P Hosseini, S Feizi

*Empirical Methods in Natural Language Processing (EMNLP), Main Conference*, 2025

### **Simulating Network Paths with Recurrent Buffering Units**

D Anshumaan\*, **S Balasubramanian\***, S Tiwari, N Natarajan, S Sellamanickam, VN Padmanabhan

*AAAI Conference on Artificial Intelligence (AAAI)*, 2023

### **Hop, Skip, and Overthink: Diagnosing Why Reasoning Models Fumble during Multi-Hop Analysis**

A Yadav, I Nalawade, S Pillarichety, Y Babu, R Ghosh, S Basu, W Zhao, W Zhao, A Nasaeh, **S Balasubramanian**, S Srinivasan

*arXiv preprint arXiv:2508.04699*, 2025

### **Seeing What's Not There: Spurious Correlation in Multimodal LLMs**

P Hosseini, S Nawathe, M Moayeri, **S Balasubramanian**, S Feizi

*arXiv preprint arXiv:2503.08884*, 2025

### **A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models**

Z Lin, S Basu, M Beigi, V Manjunatha, RA Rossi, Z Wang, Y Zhou, Y Zhou, **S Balasubramanian**, A Zarei, K Rezaei, Y Shen, B Menglong Yao, Z Xu, Q Liu, Y Zhang, Y Sun, S Liu, L Shen, H Li, S Feizi, L Huang

*arXiv preprint arXiv:2502.17516*, 2025

## Services and Teaching

---

- [SPAR](#) [🔗](#) **Mentor**: Mentoring 4 researchers (across multiple timezones) on projects related to utilizing circuits for solving AI safety problems
- Reviewer for prominent machine learning conferences such as ICML 2024, NeurIPS 2024 (Top Reviewer), ICLR 2025, NeurIPS 2025 (Top Reviewer), ICLR 2026, NeurIPS 2026.
- Introduced high-school students to AI as an instructor as part of [TRAILS AI Summer Camp](#) [🔗](#)
- Teaching Assistant for Programming Handheld Systems (CMSC 436), Probability and Statistics (STAT 400) at UMD College Park; and Data Interpretation and Analysis (CS 215) and Electricity and Magnetism (PH 108) at IIT Bombay.

## Awards and Honors

---

- Awarded Dean's Fellowship at the University of Maryland for the first year of PhD [2023]
- Awarded Institute Academic Prize for exceptional academic performance in IIT Bombay [2017]
- Ranked **2nd** in the institute out of about 900 students in the first year at IIT Bombay [2017]
- Ranked **4th** in JEE Mains out of 1.2 million candidates all over India [2016]
- Ranked **2nd** in the Maharashtra State Board Examinations (12th grade) [2016]