

Vinu Sankar Sadasivan

Final year PhD candidate
Department of Computer Science
The University of Maryland, College Park
Research Interests — AI/ML Safety


vinusankars.github.io
vinu@umd.edu
Google Scholar

EDUCATION

The University of Maryland, College Park *Aug '21 – May '25 (Expected)*
Ph.D. & M.S. in CS advised by [Prof. Soheil Feizi](#) GPA - 4.00/4.00

Indian Institute of Technology, Gandhinagar *Jul '16 – Jul '20*
B. Tech. in CSE [ [Director's Silver Medalist](#)] GPA - 9.21/10.00

INVITED TALKS

MLOps Podcast – Red-teaming for AI  *Dec '24*

UK AI Safety Institute – How to Jailbreak AI Efficiently? *Nov '24*

US Securities and Exchange Commission – Can AI-Generated Content be Reliably Detected? *May '24*

Amazon AWS Responsible AI – Fast Adversarial Attacks on Language Models *Apr '24*

Google Research – Hardness of AI Text Detection *Nov '23*

RESEARCH EXPERIENCES

Google DeepMind, Mountain View *Sep '24 – May '25*
PhD Student Researcher (full-time until Jan '25) Manager: [Dr. Lun Wang](#)

Fundamental AI Research, Meta, Paris *May – Aug '24*
Research Scientist Intern Managers: [Dr. Matthijs Douze](#), [Dr. Jakob Verbeek](#)

University of Maryland *Aug '21 – Present*
Research Assistant in CS Advisor: [Prof. Soheil Feizi](#)

IIT Gandhinagar *Aug '20 – Jul '21*
Junior Research Fellow in CSE Advisor: [Prof. Anirban Dasgupta](#)

California Institute of Technology *May – Jul '19*
Undergraduate Research Fellow in Astronomy Department Advisor: [Dr. Ashish Mahabal](#)

Microsoft Research India *Jan – Apr '19*
Research Intern in Machine Learning and Optimization Group Managers: [Dr. Harsha Simhadri](#) & [Dr. Prateek Jain](#)

Indian Institute of Science *May – Jul '17, Dec '17, Feb '18, May – Jul '18*
Research Intern at Spectrum Lab for Signal Processing Advisor: [Prof. Chandra Seelamantula](#)

RESEARCH PAPERS

* equal contribution

IconMark: Robust Interpretable Concept-Based Watermark For AI Images

VS Sadasivan, M Saberi, S Feizi

Accepted in Workshop on GenAI Watermarking 2025 at International Conference on Learning Representations (ICLR).

LLM-Check: Investigating Detection of Hallucinations in Large Language Models

G Sriramanan, S Bharti, **VS Sadasivan**, S Saha, P Kattakinda, S Feizi

Accepted at Conference on Neural Information Processing Systems (NeurIPS) 2024. [\[PDF\]](#)

DREW: Towards Robust Data Provenance by Leveraging Error-Controlled Watermarking

M Saberi, **VS Sadasivan**, A Zarei, H Mahdaviifar, S Feizi

Preprint on arXiv. June, 2024. [\[PDF\]](#)

Fast Adversarial Attacks on Language Models In One GPU Minute

VS Sadasivan, S Saha*, G Sriramanan*, P Kattakinda, A Chegini, S Feizi

Accepted at International Conference on Machine Learning (ICML) 2024. [\[PDF\]](#)

[Media Coverage](#)  [The Register](#)

Can AI-Generated Text be Reliably Detected?

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

Accepted at Transactions on Machine Learning Research (TMLR) 2025. [PDF]

[Media Coverage](#)  Nature, Washington Post, Wired, New Scientist, The Register, TechSpot

Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks

M Saberi, VS Sadasivan, K Rezaei, A Kumar, A Chegini, W Wang, S Feizi

Accepted at International Conference on Learning Representations (ICLR) 2024. [PDF]

[Media Coverage](#)  Wired, The Verge, MIT Technology Review, Bloomberg, The Register

Exploring Geometry of Blind Spots in Vision Models

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

Accepted [[spotlight](#) ☆] at Conference on Neural Information Processing Systems (NeurIPS) 2023. [PDF]

Provable Robustness for Streaming Models with a Sliding Window

A Kumar, VS Sadasivan, S Feizi

Preprint on arXiv. March, 2023. [PDF]

CUDA: Convolution-based Unlearnable Datasets

VS Sadasivan, M Soltanolkotabi, S Feizi

Accepted at Computer Vision and Pattern Recognition Conference (CVPR) 2023. [PDF]

Statistical Measures For Defining Curriculum Scoring Function

VS Sadasivan, A Dasgupta

Accepted [[spotlight](#) ☆] at SubSetML Workshop at International Conference on Machine Learning (ICML) 2021. [PDF]

Shallow RNN: Accurate Time-series Classification On Resource Constrained Device

D Dennis, DAE Acar, V Mandikal, VS Sadasivan, V Saligrama, HV Simhadri, P Jain


Accepted at Conference on Neural Information Processing Systems (NeurIPS) 2019. [PDF]


High Accuracy Patch-Level Classification Of Wireless Capsule Endoscopy Images Using A Convolutional Neural Network

VS Sadasivan, CS Seelamantula

Accepted at IEEE International Symposium on Biomedical Imaging (ISBI) 2019. [PDF]

AWARDS AND HONORS

[Kulkarni Fellowship Awardee](#)  at University of Maryland in 2023.

[Notable reviewer](#)  top ~1% reviewer in ICLR 2023.

[Director's Silver Medalist](#)  CSE, IIT Gandhinagar in 2020.

[Special mention for poster](#) Undergraduate Research Conclave, IIT Gandhinagar in 2019.

[Summer Undergraduate Research Fellowship](#)  Caltech in 2019 (awarded ~ 6,350 USD).

[Kerala State Topper, Regional Mathematics Olympiad](#) in 2014.

[KVPY awardee](#) by Government of India in 2016. Ranked 85 out of ~ 100,000 in the country.

[NTSE scholar](#) awarded by Government of India in 2012.

SERVICES & TEACHING

Reviewer for prominent machine learning conferences such as ICML 2021, NeurIPS 2022, ICLR 2023 ([Notable reviewer](#)), NeurIPS 2023, ICML Neural Compression Workshop 2023, ICML 2024, TACL, ICML 2025, ICCV 2025.

Teaching assistant for CMSC720: Foundations of Deep Learning (Spring 2024), CMSC 422: Introduction to Machine Learning (Fall 2021), and CMSC 320: Introduction to Data Science (Spring 2022) at UMD.

Peer-assisted learning mentor at IIT Gandhinagar, helping freshmen who found it difficult to cope with their academic workload.