Vedant Nanda

✉ nvedant07@gmail.com | ⌂ nvedant07.github.io | nvedant07 | @_nvedant_

Work Experience

Aleph Alpha
Heidelberg, DE

AI Researcher — *June 2024 - Present*

Fast and controllable LLM inference: We build our custom LLM serving platform that achieves high overall throughput using efficient KV Cache management (eg: PagedAttention) and provides novel inference-time capabilities such as steering and explainability.

Amazon AWS
WA, USA

Applied Science Intern — *June 2023 - August 2023*

AWS Bedrock: I built an inference time de-biasing algorithm for text-to-image diffusion models.

Amazon AWS
Cambridge, UK

Applied Science Intern — *November 2022 - January 2023*

AWS Clarify: I built an actionable causal explainability method using manifold constraints that outperformed existing methods for the feasibility of acting upon the explanations and could be applied to a broad family of ML models on tabular data.

Education

University of Maryland, College Park
MD, USA & SB, DE

Ph.D. in Computer Science — *August 2019 - May 2024*

- Research Interests: Trustworthy Deep Learning: Fairness, Robustness, and Efficiency
- Part of Maryland-Max Planck joint program through which I spent time at MPI-SWS in Saarbrücken, DE.
- Advisors: Krishna P. Gummadi (MPI-SWS) and John P. Dickerson (University of Maryland)

University of Maryland, College Park
College Park, MD, USA

M.S. in Computer Science — *August 2019 - May 2022*

- Advisor: John P. Dickerson
- TA for CMSC 320 (Intro to Data Science), Fall 2019

Indraprastha Institute of Information Technology (IIIT) Delhi
New Delhi, India

B.Tech. in Computer Science and Engineering — *August 2015 - May 2019*

- **GPA: 9.47/10**, in top 5% of institute
- Part of dean's list for academic excellence for all years
- Selected Coursework: Numerical Methods, Calculus-I, Calculus-II, Machine Learning, Collaborative Filtering, Information Retrieval, Designing Human-Centered Systems, System Administration, Network Administration

Publications

The Impact of Inference Acceleration Strategies on Bias of Large Language Models
NAACL 2025

Elisabeth Kirsten, Ivan Habernal, **Vedant Nanda**, Muhammad Bilal Zafar

An earlier version appeared at SafeGenAI workshop at NeurIPS 2024

Lawma: The Power of Specialization for Legal Tasks
ICLR 2025

Ricardo Dominguez-Olmedo, **Vedant Nanda**, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, Michael Livermore

An earlier version appeared at Symposium on CS&Law

Towards Reliable Latent Knowledge Estimation in LLMs: Zero-Prompt Many-Shot Based Factual Knowledge Extraction
WSDM 2025

Qinyuan Wu, Mohammad Aflah Khan, Soumi Das, **Vedant Nanda**, Bishwamittra Ghosh, Camila Kolling, Till Speicher, Laurent Bindschaedler, Krishna P. Gummadi, Evimaria Terzi

Code: https://github.com/QinyuanWu0710/ZeroPrompt_LKE, HF Dataset: https://huggingface.co/datasets/QinyuanWu/T-Rex-MC

Understanding the Role of Invariance in Transfer Learning
TMLR 2024

Till Speicher, **Vedant Nanda**, Krishna P. Gummadi

Code: github.com/tillspeicher/representation-invariance-transfer

Diffused Redundancy in Pre-trained Representations
NeurIPS
Vedant Nanda, Till Speicher, John P. Dickerson, Krishna P. Gummadi, Soheil Feizi, Adrian Weller
2023
Code: github.com/nvedant07/diffused-redundancy

What Happens During Finetuning of Vision Transformers: An Invariance Based Investigation
Conference on Lifelong Learning Agents (CoLLAs)
Gabriele Merlin, **Vedant Nanda**, Ruchit Rawal, Mariya Toneva
2023

Do Invariances in Deep Neural Networks Align with Human Perception?
AAAI (Oral)
Vedant Nanda, Ayan Majumdar, Camila Kolling, John P. Dickerson, Krishna P. Gummadi, Bradley C. Love, Adrian Weller
2023
Code: github.com/nvedant07/Human-NN-Alignment

Rawlsian Fairness in Online Bipartite Matching: Two-sided, Group, and Individual
AAAI
Seyed A. Esmaeili, Sharmila Duppala, Davidson Cheng, **Vedant Nanda**, Aravind Srinivasan, John P. Dickerson
2023
Earlier version appeared as extended abstract at AAMAS 2022

Measuring Representational Robustness of Neural Networks Through Shared Invariances
ICML (Long Oral)
Vedant Nanda, Till Speicher, Camila Kolling, John P. Dickerson, Krishna P. Gummadi, Adrian Weller
2022
Code: github.com/nvedant07/STIR

Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning
FAccT
Vedant Nanda*, Samuel Dooley*, Sahil Singla, Soheil Feizi, John P. Dickerson
2021
* Equal Contribution; Code: github.com/nvedant07/Fairness-Through-Robustness

Balancing the Tradeoff between Profit and Fairness in Rideshare Platforms during High-Demand Hours
AAAI
Vedant Nanda, Pan Xu, Karthik A. Sankararaman, John P. Dickerson, Aravind Srinivasan
2020
Also presented at AIES 2020 (Oral); Code: github.com/nvedant07/rideshare-fairness-peak

On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning
ICML
Hoda Heidari *, **Vedant Nanda ***, Krishna P. Gummadi
2019
* Equal Contribution; Code: github.com/nvedant07/effort_reward_fairness

Leveraging Facebook's Free Basics Engine for Web Service Deployment in Developing Regions
ICTD
Siddharth Singh*, **Vedant Nanda***, Rijurekha Sen, Satadal Sengupta, Ponnurangam Kumaraguru, Krishna P. Gummadi
2017
* Equal Contribution

Workshops and Posters

Learning to Explain Machine Learning
CHI workshop on Human-Centered Explainable AI
Vedant Nanda*, Duncan McElfresh*, John P. Dickerson
2021
* Equal Contribution

Technical Challenges for Training Fair Neural Networks
ICLR workshop on Responsible AI
Valeriia Cherepanova*, **Vedant Nanda***, Micah Goldblum, John P. Dickerson, Tom Goldstein
2021
* Equal Contribution

Unifying Model Explainability and Robustness via Reasoning Labels
NeurIPS workshop on Safety and Robustness in Decision Making
Vedant Nanda, Junaid Ali, Krishna P. Gummadi, Muhammad Bilal Zafar
2019

Stop the KillFies! Using Deep Learning Models to Identify Dangerous Selfies
WWW workshop on Modelling Social Media
Vedant Nanda, H.Lamba, D.Agarwal, M.Arora, N.Sachdeva, P.Kumaraguru
2018

Empirical Analysis of Facebook's Free Basics *SIGMETRICS (poster)*

S.Singh*, **Vedant Nanda***, R.Sen, S.Ahmad, S.Sengupta, A.Phokeer, Z.A.Farooq, T.A.Khan, P.Kumaraguru, I.A.Qazi, D.Choffnes, K.P.Gummadi *2017*

* Equal Contribution

Other Experience

University of Maryland, College Park *MD, USA*

Teaching Assistant, CMSC320: Intro to Data Science *Aug 2019 - Dec 2019*

Advisor: John P. Dickerson

Max Planck Institute for Software Systems *Saarbrücken, Germany*

Research Intern *May 2018 - Aug 2018*

Advisor: Krishna P. Gummadi

Precog, IIITD *New Delhi, India*

Research Intern *May 2017 - Aug 2017*

Advisor: Ponnurangam Kumaraguru

Presentations/Talks

2024	Talk @ Sysnets MPI-SWS on Efficient and Controllable LLM Inference. Hosted by Laurent Bindschaedler.
2024	Talk @ AI Fest on Efficient and Controllable LLM Inference. Hosted by Arthur AI.
2024	Talk @ Ruhr Universität Bochum on Efficient and Controllable LLM Inference. Hosted by Muhammad Bilal Zafar.
2024	Talk @ ETH Center for Law and Economics on Specializing LLMs for Legal Tasks. Hosted by Stefan Bechtold.
2024	Thesis Defense at University of Maryland.
2024	Talk @ Huawei Research "Towards Foundations of Trustworthy Deep Learning: Fairness, Robustness and Efficiency".
2024	Talk @ Bosch Center for AI "Towards Foundations of Trustworthy Deep Learning: Fairness, Robustness and Efficiency".
2023	Thesis Proposal at University of Maryland.
2022	Talk at University of Cambridge Machine Learning Group. Hosted by Adrian Weller.
2022	Oral Talk at International Conference on Machine Learning (ICML), Baltimore, Maryland.
2022	Talk at Computer Vision and Machine Learning seminar @ MPI-INF, virtual.
2022	Talk at ML Tea @ MPI-SWS, virtual.
2021	Talk at UMD Fairness in AI Seminar, *joint with Valeriia Cherepanova*, virtual. Link.
2021	Paper QnA at Conference on Fairness Accountability and Transparency (FAccT), virtual. Link.
2020	Oral talk at Conference on AI, Ethics and Society (AIES), NYC, USA

Service

Reviewer	ICML 2021, 2023, 2024, 2025
	ICLR 2023, 2025
	NeurIPS 2021
	AAAI 2021
	CVPR 2021
	ICCV 2021
	WWW 2020, 2021
	ASONAM 2019
Other	UMD Graduate Admission Reviewer 2020
	ELLIS PhD Admission Reviewer 2023

Skills

ML	PyTorch, Lightning/LitGPT, Transformers, Accelerate, Numpy, Triton/CUDA, DeepSpeed
Other	Matplotlib, Pandas, Git, C/C++

References

1. **Prof. Krishna P. Gummadi**
   Scientific Director
   Max Planck Institute for Software Systems
2. **Prof. John P. Dickerson**
   Associate Professor, Computer Science
   University of Maryland, College Park