Neel Sindhar Jain

njain17@umd.edu

Google Scholar | Github | LinkedIn | Website

Education

2023 – present Ph.D., Computer Science, **University of Maryland, College Park** (Advisor: Tom Goldstein)
2021 – 2023 M.S., Computer Science, **University of Maryland, College Park**
2016 – 2020 B.A., Honors in Mathematics, **Williams College**

Research Overview

The widespread adoption of LLMs highlights the pressing need to ensure these systems are both safe and trustworthy. Thus, Neel's research focuses on Safe and Trustworthy Machine Learning, where he explores deeply interconnected challenges spanning refusal behavior, safety, evaluations, and reasoning.

Citations: 1.5k+ Github Stars: 500+

Selected Work

DynaGuard: A Dynamic Guardrail Model With User-Defined Policies.
Monte Hoover*, Vatsal Baherwani, **Neel Jain**, Khalid Saifullah, Joseph Vincent, Chirag Jain, Melissa Kazemi Rad, C. Bayan Bruss, Ashwinee Panda, Tom Goldstein. *arXiv:2509.02563*, 2025. Link

Scaling up test-time compute with latent reasoning: A recurrent depth approach.
Jonas Geiping*, Sean McLeish, **Neel Jain**, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Tom Goldstein. *Neurips (Spotlight)*, 2025. Link

Refusal Tokens: A Simple Way to Calibrate Refusals in Large Language Models.
Neel Jain*, Aditya Shrivastava, Chenyang Zhu, Daben Liu, Alfy Samuel, Ashwinee Panda, Anoop Kumar, Micah Goldblum, Jonas Geiping, Tom Goldstein. *Conference on Language Modeling (COLM)*, 2025. Link

LiveBench: A Challenging, Contamination-Limited LLM Benchmark.
Colin White*, Samuel Dooley*, Manley Roberts*, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, **Neel Jain**, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, Micah Goldblum. *International Conference on Representation Learning ICLR (Spotlight), 2025*. Link

NEFTune: Noisy Embeddings Improve Instruction Finetuning.
Neel Jain*, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, Tom Goldstein. *International Conference on Representation Learning (ICLR)*, 2024. Link

Bring Your Own Data! Self-Supervised Evaluation for Large Language Models.
Neel Jain*, Khalid Saifullah*, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, Tom Goldstein. *Conference on Language Modeling (COLM)*, 2024. Link

Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery.
Yuxin Wen*, **Neel Jain***, John Kirchenbauer, Micah Goldblum, Jonas Geiping, Tom Goldstein. *NeurIPS*, 2024. Link

Baseline Defenses for Adversarial Attacks Against Aligned Language Models.
Neel Jain*, John Kirchenbauer, Paras Jain, Jonas Geiping, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Tom Goldstein. *arXiv:2309.00614*, 2023. Over 500 citations. Link

Work Experience

Residence Builders Program, Mozilla.ai Jun 2025 – Present

Advised by Prof. Tom Goldstein & John Dickerson. Investigating gradient-based methods for datafiltering in Midtraining.

Graduate Researcher, University of Maryland, College Park Various terms from Jun 2022

Advised by Prof. Tom Goldstein. LLM safety/guardrails (PEZ optimizer/Refusal Messages/Guardian Models), contamination-limited evaluation (BYOD/LiveBench), instruction finetuning (NEFTune).

Applied Research Intern, Capital One Jun 2024 – Aug 2024

Finetuning LLM Team: investigated how to control refusal messages in LLMs.

Teaching Assistant (combined), UMD & Williams College Various terms till May 2022

Outstanding Graduate Teaching Assistant Award Recipient in Fall of 2021. UMD: multiple CS/ML courses (discussion, office hours, grading, project support). Williams: *Introduction to Mechanics* (Prof. Willimas Wootters). *Fun fact: Prof. Wooters helped coin the term qubit.*

Data Scientist, Booz Allen Hamilton Jul 2019 – Apr 2021

Built analytics models (agent-based, Monte Carlo), production web apps (Flask), and data pipelines; supported programs incl. DoD OSD CAPE.

Research Intern, Salk Institute for Biological Studies May 2017 – Aug 2017

Computational biology in the Edward Stites Lab; data analysis and modeling using ODEs.

Skills

ML/LLMs	Instruction finetuning, guardrails/safety, evaluation/benchmarking, distributed training
Systems	PyTorch, Transformers, Large Scale Distributed Training with AMD