Jaydeep Jitendra Borkar

jaijborkar@gmail.com
http://jaydeepborkar.github.io/
https://github.com/jaydeepborkar

Education & Research Experience

Northeastern University *Sept 2021 - present*
PhD in Computer Sciences
Advisor: Prof. David Smith

MIT-IBM Watson AI Lab *August 2020 - July 2021*
External Research Student
Advisor: Dr. Pin-Yu Chen
Worked on developing new and simple methods for adversarial image generation that fool real-world vision APIs.

Savitribai Phule Pune University *2016 - 2020*
Bachelor's degree in Computer Engineering

CIFAR Deep Learning + Reinforcement Learning Summer School *Aug 2020*
Hosted by Mila (*25%* acceptance rate)
Amongst 300 students selected across 45 countries for the summer school

Research Interests: NLP privacy and safety, memorization in LLMs, Generative AI safety.

Skills: Python, PyTorch, Transformers, fine-tuning large language models, Numpy, Hugging Face libraries, R, Pandas, Matplotlib, CUDA.

Papers & Research Projects

What can we learn from Data Leakage and Unlearning for Law?
Jaydeep Borkar
ICML 2023 Generative AI and Law (GenLaw) workshop, Honolulu, Hawaii
Link: https://genlaw.github.io/CameraReady/12.pdf

Semantic Memorization in LLMs *ongoing*
Mentors: Naomi Saphra and Katherine Lee
Working on categorizing different types of memorization in Pythia models and analyzing attention patterns for memorized and non-memorized examples.

Interpretability in Transformer OCR Models *ongoing*
Advisor: Prof. David Smith
Working on interpretability-related research for transformer-based optical character recognition models such as trocr.

Extracting Training Data from Pre-trained and Fine-tuned GPT-2
CS 7150 Deep Learning class project
Jaydeep Borkar
Showed that fine-tuned models can memorize and leak both fine-tuning and pre-training data during text generation. Project report: `https://jaydeepborkar.github.io/7150_project_report.pdf`

Simple Transparent Adversarial Examples
Jaydeep Borkar and Pin-Yu Chen
ICLR 2021 Workshop on Security and Safety in Machine Learning Systems
Link: https://aisecure-workshop.github.io/aml-iclr2021/papers/48.pdf

Organizing

Trustworthy ML Initiative
Co-organizer of the Trustworthy ML Initiative along with Prof. Hima Lakkaraju (Harvard), Sara Hooker (Cohere

for AI), Dr. Sarah Tan, Dr. Subho Majumdar, Chhavi Yadav (UC San Diego), Dr. Chirag Agarwal (Harvard), Prof. Haohan Wang (UIUC), and Marta Lemanczyk (Hasso-Plattner-Institut).

Courses

Machine Learning CS 6140, Natural Language Processing CS 6120, Deep Learning CS 7150, Machine Learning Security and Privacy CY 7790, Theory and Methods in Human-Computer Interaction CS 7340.

Teaching Experience

Foundations of AI CS 5100- TA	*Spring 2024*
Foundations of Data Science DS 3000 - TA	*Fall 2023*
Product Development for Large Language Models CS 7180 - TA	*Summer 2023*
Introduction to Computer Science Research CS 3950 and CS 4950 - TA	*Spring 2023*
Introduction to Machine Learning and Data Mining DA 5030 - TA	*Summer and Fall 2022*

Awards and Honors

- Travel Grant Award to attend the first IEEE conference on Secure and Trustworthy Machine Learning (**SaTML**). 2022
- ICML 2021 Travel Grant Award for Safety and Security in Machine Learning Systems workshop. 2021
- Accepted to CIFAR Deep Learning + Reinforcement Learning Summer School. Amongst **300** students selected across **45** countries. 2020
- Awarded student grant to attend USENIX Security 2020 2020
- Poster speaker at PyCon India. 2019