

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 A. 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: Privacy and safety in language models, training data extraction (memorization) in LLMs, Generative AI safety.

Skills: Python, PyTorch, Transformers, Numpy, Hugging Face, Pandas, 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>

Memorization of PII in Language Models

working paper

Jaydeep Borkar, Christopher A. Choquette-Choo, Katherine Lee, Matthew Jagielski, David A. Smith

Studying leakage of sensitive information such as PII

Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, **Jaydeep Borkar**, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, Naomi Saphra

Preprint: <https://arxiv.org/abs/2406.17746>

Mind the gap: Analyzing lacunae with transformer-based transcription

Jaydeep Borkar and David A. Smith

ICDAR 2024 Workshop on Computational Paleography

Link: <https://arxiv.org/abs/2407.00250>

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.

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 (Salesforce AI), Dr. Subho Majumdar (Vijil), 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, AI as an Archival Science CS 7180.

TEACHING EXPERIENCE

Natural Language Processing CS 6120 - TA	<i>Summer 2024</i>
Foundations of AI CS 5100- TA	<i>Spring 2024</i>
Foundations of Data Science DS 3000 - TA	<i>Fall 2023</i>
Product Development for Large Language Models CS 7180 - TA	<i>Summer 2023</i>
Introduction to Computer Science Research CS 3950 and CS 4950 - TA	<i>Spring 2023</i>
Introduction to Machine Learning and Data Mining DA 5030 - TA	<i>Summer and Fall 2022</i>