Sana Pandey

Dynamic, driven, and organized with experience in building human-oriented AI and technology policy.

sanapnde@mit.edu sanapandey.github.io sanapandey

Education

Massachusetts Institute of Technology, Graduate Student in EECS and Technology Policy 2025-2027

- **Research:** Currently at CSAIL at the Language and Intelligence Group under Jacob Andreas. Fully supported by the National Science Foundation's Graduate Research Fellowship.
- **Relevant Coursework:** Reinforcement Learning, Economic Analysis for Business Decisions.

University of California, Berkeley, Bachelor's in Data Science and Cognitive Science 2024

- 3.88/4.0 GPA. Minored in Mandarin Chinese. Captain of Women's Epee Fencing Team.
- **Relevant Coursework:** Artificial Intelligence; Modeling, Learning, and Decision Making; Data Structures; Structure and Interpretation of Computer Programs; Linear Algebra; Discrete Mathematics and Probability; Designing Algorithmic Media (Audited Graduate Course).

Stanford University, China Scholars Program 2020

- **Submitted Thesis:** "Rice Rabbit: Analyzing the Evolution of China's Feminist Movement with the Rise of Censorship and Social Media"

The Harker School, High School Diploma 2020

- National Merit Scholar Finalist (2020). AP Scholar with Distinction (2020). Mission of the School Award (2019). Presidential Service Award (2019, 2020).

Industry Experience

Research Engineer, Applied AI and Recommender Systems, Center for Human-Compatible AI Fall 2024 - Summer 2025

- **Developing** LLM and recommendation model evaluations grounded in game theory and social choice theory. Building white-box evaluations for closed-source models towards explainability research. Applying statistical methods to model optimization and training.

Chief Technology Officer, Hortus AI Fall 2024

- **Built** out AI assessment platforms grounded in human feedback with a team of two. Developed scalable MVPs grounded in client feedback, working with the City of Boston to develop chatbot metrics for reinforcement learning systems reaching over 650k residents.

Analytics and Machine Learning Intern, Apple Winter 2023

- **Implemented** a graph neural network (Node2Vec, Tensorflow Keras), generating context-driven next-step recommendations through second-order random walks in an integrated application.
- **Integrated** a user interface (Streamlit) that updated predictions based on adjustable parameters in real time.

Machine Learning and AI Intern, Woebot Inc. Summer 2022

- **Spearheaded** the creation and implementation of topic modeling using DistilBERT, BERTopic, HDBScan, and NLTK to run analysis on user messaging input and ensure content relevance, creating a 23% improvement in user experience.
- **Innovated** new functionalities tracking topic emergence/growth, built software to automatically alert developers of new classifier categories, and retrained existing classifier models to over 90% precision and recall in all user content domains.

Synthetic Biology and Machine Learning Intern, Koniku Inc. Fall 2020

- **Project-lead** for Covid-19 research project, tracking product development and isolating 20+ receptor protein constructs to detect virus presence. Modeled impact of oxidative stress on neuronal cells across 200+ gene constructs.

Research Experience

Graduate Research Assistant, Language and Intelligence Group (LINGO), MIT CSAIL — Fall 2025-Present

- **Researching** how data features shape internal knowledge representations in models, and testing downstream generalization effects
- **Constructing** long-term optimization metrics that encode human values in applied settings

Research Intern, Center for Human-Compatible AI at UC Berkeley — Spring 2022 – Summer 2024

- **Prosocial Ranking Challenge:** Built out infrastructure to test multiple implementations of prosocial recommendation algorithms on social media platforms. Implemented example code for natural language processing, statistical, and LLM-driven rankers for challenge participants.
- **Clinical Diagnostics Model:** Using Node2Vec and Tensorflow, built a sequence-to-sequence prediction model designed to streamline session flow and treatment plans for admitted patients. Innovated new methodology to preserve probabilities of diagnoses post-tokenization. Developed in collaboration with the **UCSF Center for Computational and Precision Health**.
- **Non-Engagement Con:** Co-organized industry-academia workshop to discuss proprietary experiment results of ranking interventions and feedback integration in platform systems. Co-authored *What We Know About Non-Engagement Signals in Content Ranking.*
- **LLM Jailbreak Benchmark**: Individually evaluated eight jailbreaking strategies, and generated categories of high-risk output for language model testing. Co-authored *A StrongREJECT for Empty Jailbreaks.*

Science and Technology Policy Intern, The Lauder Institute at the University of Pennsylvania — Fall 2022

- **Analyzed** science and technology policy based on topic frequency and input, curating data on over 150 think tanks and 100 academic sources to the Global Go To Think Tank Database.
- **Featured** in the World Economic Forum and United Nations Center for Development.

Neural Modeling Intern, The CognAc Lab at UC Berkeley — Fall 2021

- **Modeled** implicit learning processes within sensorimotor adaptation under Prof. Rich Ivry, and extended findings towards machine learning generalization.

Technology Policy Intern, Blum Center for Developing Economies at UC Berkeley — Spring 2021

- **Researched** intellectual property law and open-source software development through the lens of international cultural influence with Prof. Clare Talwalker.

Publications

What We Know About Non-Engagement Signals in Content Ranking — 2023
Tom Cunningham, **Sana Pandey**, Leif Sigerson, Jonathan Stray, Jeff Allen, Bonnie Barrilleaux, Ravi Iyer, Smitha Milli, Mohit Kothari, Behnam Rezaei.
Currently an invited submission at the New York Academy of Sciences Special Issue. Available on arXiv here: https://arxiv.org/abs/2402.06831

A StrongREJECT for Empty Jailbreaks — 2024
Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, **Sana Pandey**, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, Sam Toyer.
Accepted into NeurIPS and ICLR. Available on arXiv here: https://arxiv.org/abs/2402.10260

Constructive Dialogue or Chaos? Assessing Online Content via Comment Interactions — 2024
Sana Pandey, Juan Leviano, Jonathan Stray.
Pre-print available soon.

Awards and Honors

National Science Foundation Graduate Research Fellowship — 2025
Graduation with Distinction — 2024
USA Collegiate Fencing Nationals, Top 10 — 2024
UC Berkeley Dean's List — 2021, 2023

Junior Olympian, Fencing 2019, 2020
National Security Language Initiative for Youth Scholarship–Taipei, Taiwan 2019

Featured Talks

Cooperation and Coordination Panel Host at CHAI Annual Workshop 2025
Invited Speaker at UC Berkeley Haas School of Business 2022-2024
Societal Effects of AI Panel Host at CHAI Annual Workshop 2024

Media Appearances

UC Berkeley Department of Computing, Data Science, and Statistics 2024

- Article title: Sana Pandey uses AI to shape a brighter future for society.

CBS Evening News 2023

- Feature title: Computer science student at UC Berkeley develops tech to combat social media harms.

National Society of Women Engineers 2021

- Podcast Episode Title: Gender Equity in STEM.