Sana Pandey | sanapandey@berkeley.edu |

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

Educational History

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

Minored in Mandarin (Chinese). Studied Abroad at the University of Padova in 2023.

Captain of the Women's Epee Fencing Team. Individually ranked in top 10 at USA Collegiate Fencing Nationals.

Relevant Coursework: Artificial Intelligence; Modeling, Learning, and Decision Making; Data Structures; Structure and Interpretation of Computer Programs; Linear Algebra; Discrete Mathematics and Probability.

Stanford University – China Scholars Program *2020*

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*

Leadership Award (2020). Mission of the School Award (2019). Love of Learning Award (2017, 2018, 2019, 2020).

Industry Experience

Chief Technology Officer and AI Engineer, Hortus AI *Fall 2024-Present*

Building out dynamic feedback and evaluation platforms for public-facing AI systems developed by government.

Analytics and Machine Learning Intern, Apple Inc. *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.

Created a user interface (Streamlit) that updated predictions based on adjustable parameters in real time, and integrated multiple models into the application for deployment.

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 more than 20 receptor protein constructs to detect virus presence. Modeled impact of oxidative stress on neuronal cells across more than 200 gene constructs.

Research Experience

Research Intern, Center for Human-Compatible AI (CHAI) *Spring 2022-Present*

Researching recommender systems, LLM evaluation, and technology policy with Jonathan Stray and Stuart Russell.

Science and Technology Policy Intern, University of Pennsylvania's Lauder Institute *Fall 2022*

Conducted quantitative analysis on trends in 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 World Economic Forum.

Neural Modeling Intern, UC Berkeley. *Fall 2021*

Modeled and researched implicit learning processes and motor behavior within sensorimotor adaptation under Prof. Rich Ivry.

Technology Policy Intern, 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.

Abstract Excerpt: We believe that there is much unrealized potential in including non-engagement signals, which can improve outcomes both for platforms and for society as a whole. Based on a daylong workshop with experts from industry and academia, we formulate a series of propositions and document each as best we can from public evidence, including quantitative results where possible.

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.

Abstract Excerpt: Most jailbreak papers claim the jailbreaks they propose are highly effective, often boasting near-100% attack success rates. However, it is perhaps more common than not for jailbreak developers to substantially exaggerate the effectiveness of their jailbreaks. We suggest this problem arises because jailbreak researchers lack a standard, high-quality benchmark for evaluating jailbreak performance, leaving researchers to create their own. We show that existing benchmarks suffer from significant shortcomings and introduce the StrongREJECT benchmark to address these issues.

Constructive Dialogue or Chaos? Assessing Online Content via Comment Interactions *In Progress*

Sana Pandey, Juan Leviano, Jonathan Stray.

Preprint available soon, planned for 2025 submission.

Awards and Honors

Graduation with Distinction *2024*

Awarded by UC Berkeley to the top 10% of students across all colleges in a graduating class.

UC Berkeley Deans' List *2021, 2022, 2023*

Awarded by UC Berkeley to the top 10% of students each semester for each college.

National Merit Scholarship Finalist *2020*

Awarded by the College Board to students scoring in the top 1% nationwide in the SAT/ACT.

Junior Olympian, Fencing *2019, 2020*

Qualified in the top 10% of all fencers in Women's Epee nationally.

National Security Language Initiative for Youth Scholarship–Taipei, Taiwan *2019*

Awarded by the Department of State to highly qualified students for language immersion and policy study.

Featured Talks

Recurring Speaker at UC Berkeley's Haas School of Business *2022-Present*

Course Titles: Ethics and AI, AI and the Future of Business

Panel Host at the CHAI Annual Workshop *2024*

Session Title: Societal Effects of AI

Media Appearances

Featured by UC Berkeley's Department of Computing, Data Science, and Statistics *2024*

Article Title: Sana Pandey uses AI to shape a brighter future for society

Featured on CBS News *2023*

Article Title: Computer science student at UC Berkeley develops tech to combat social media harms

Featured on National Society of Women Engineers' Advocacy Podcast *2021*

Episode Title: Gender Equity in STEM

Projects

AI Development Tracker, Center for Human-Compatible AI *2024*

Using HTML, JavaScript, Langchain, and Figma, built a page that scrapes, categorizes, and summarizes technical developments in AI innovation. Designed to make AI developments accessible and interpretable to the public.

Prosocial Ranking Challenge, Center for Human-Compatible AI *2024*

In collaboration with a team of researchers, built out infrastructure to test multiple implementations of prosocial recommender algorithms on social media platform software.

Clinical AI Diagnostics Project, UCSF Center for Computational Precision Health *2023*

Using Node2Vec and Tensorflow, built a sequence-to-sequence recommender model designed to streamline session flow and treatment plans for admitted patients and divergence from optimal path through an actor-critic framework.

Cut sections:

Relevant Skills

Coding Languages [In Order of Proficiency]: Python, Java, SQL, C++, JavaScript, HTML, MatLAB, Scheme, Node.js, Ruby on Rails.

Libraries: Tensorflow, PyTorch, Transformers (Huggingface), Pandas, Matplotlib, Seaborn, DistilBERT, NLTK, Word2Vec, Node2Vec, HDBScan, spaCy, Langchain, Numpy.

Spoken Languages [In Order of Proficiency]: English, Hindi, Mandarin, Korean, Italian, American Sign Language.

Longer Awards and Honors

Graduation with Distinction *2024*

Awarded by UC Berkeley to the top 10% of students across all colleges in a graduating class.

UC Berkeley Deans' List *2021, 2022, 2023*

Awarded by UC Berkeley to the top 10% of students each semester for each college.

National Merit Scholarship Finalist *2020*

Awarded by the College Board to students scoring in the top 1% nationwide in the SAT/ACT.

AP Scholar with Distinction *2020*

Awarded by the College Board to students with an average score >4 over 8+ AP exams, traditionally the top 1%.

Junior Olympian, Fencing

Qualified in the top 10% of all fencers in Women's Epee nationally. *2019, 2020*

Presidential Service Award *2019, 2020*

Awarded to high school students exhibiting exemplary dedication to community service.

Projects:

Image Recog *2022*

Using Tensorflow, HTML, and JavaScript, coded a web application that runs image recognition on uploaded pictures and isolates relevant popular phases and hashtags.