Shreya Kochar

shreyako@cs.cmu.edu | linkedin.com/in/shreyakochar/ | github.com/skochar1/

Education

Columbia University GPA: 4.099/4.00
Masters of Science in Computer Science *Jan. 2023 – May 2025*

Wellesley College GPA: 3.86/4.00
Bachelor of Arts in Computer Science - Magna Cum Laude *Aug. 2019 – May 2022*

Massachusetts Institute of Technology GPA: 5.00/5.00
Cross Registered Student *Aug. 2019 – May 2022*

Experience

Pre-doctoral Research Fellow (Project Scientist) Oct. 2025 – Present
Carnegie Mellon University, School of Computer Science *Pittsburgh, PA*

- Appointed by Professor Norman Sadeh to join the *Smart City Privacy Technologies* project as a Project Scientist with postdoctoral-level responsibilities.
- Designing and deploying new privacy-preserving and AI-enabled features for Carnegie Mellon's IoT Privacy Infrastructure, supporting 100,000+ IoT resource descriptions and tens of thousands of end users.
- Conducting empirical evaluation with the City of Long Beach government, integrating human-centered design with privacy engineering to advance civic technology.
- Collaborating with interdisciplinary teams and government partners

Pre-doctoral Research Fellow Jan. 2025 – Present
University of Chicago, Department of Computer Science *Chicago, IL*

- Selected as a Predoctoral Research Fellow to work with Professor Nick Feamster on privacy and safety questions surrounding large language models.
- Investigating privacy risks and bias in LLMs through survey methodology and applied statistical analysis.
- Co-authored a Google grant proposal, *Understanding the Privacy and Safety Risks of Mental-Health Chatbots*, designing methods to benchmark therapeutic quality and audit privacy policies of mental-health AI tools
- Working on project in collaboration with Stanford to evaluate LLM vs. human perception of privacy harms.

Founding Engineer (ML / LLM Engineering) Aug. 2025 – Oct. 2025
Avenio Corporation *San Francisco, CA*

- Built core infrastructure for **Avenio.ai**, a generative-AI platform for bettering clinical trials, using Python, Django, Celery, Qdrant, and LangChain.
- Designed and implemented ingestion and vector-search pipelines (ETL, embeddings, caching layers) for multimodal biomedical content, enabling sub-second retrieval.
- Built and optimized ingestion + retrieval pipelines for biomedical content (PubMed, ClinicalTrials.gov, FDA, etc.) to reduce hallucination rates and improve answer reliability.
- Developed medical-source ranking and URL-validation services (40+ domains, parallel checks, caching), ensuring 100% of returned references are live and authoritative.

Software Engineer/Data Engineer May 2024 – Aug 2025
Microsoft C + AI: Audits and Risks Team *Redmond, WA*

- Integrated quicker Azure OpenAI querying into our backend processes (increasing the query speed from 20 seconds to 0.5 seconds), enabling proactive audit data analysis and insights generation
- Designed and implemented a scalable Data Quality Framework leveraging Azure Synapse Analytics, Azure Data Lake Storage, and parameterized notebooks, enabling automated daily partitioning and trend analysis of data quality results for seamless PowerBI integration.
- Converted legacy financial data processes to automated daily refreshes in Synapse, delivering time savings of 3 weeks per quarter and enhancing accuracy for business reporting.

Computer Science Teaching Assistant Aug. 2024 – May 2025
Columbia Computer Science Department *Morningside Heights, NY*

- Worked as a TA for the Advanced Software Engineering course (COMS W4156) in Fall 2024.
- Worked as a TA for the Topics in Software Engineering research course (COMS E6156) in Spring 2025.
- Helped create the course curriculum, homework assignments, and exam questions. Lecture in class/demo skills to use for final project and homework assignments.

Research Assistant – Predictive Privacy Project Jan 2023 – May 2025
Columbia University, Computer Science Department *New York, NY*

- Collaborated with Professor Steven M. Bellovin to design and implement the Predictive Privacy framework, an empirical method to quantify privacy harms in data-sharing and inference scenarios.
- Co-authored law-review and technical papers. Project details below.
- Supervised an undergraduate assistant and managed end-to-end experiments, from IRB approval to data analysis.

Software Engineer Nov. 2022 – May 2024
Microsoft C and AI: Security Team *Redmond, WA*

- Created a template service to help teams transition from using less secure authorization methods (certificates/secrets) to identities; reduced the time required for identity integration/adoption from months to a couple of weeks
- Used logs to implement active learning and output the role(s) that uses the least amount of privileges for a given task

Software Engineering Intern May 2022 – Aug. 2022
Microsoft *Redmond, WA*

- Machine Learning intern at Microsoft's Commerce and Ecosystems department
- Trained, tested, and cross validated several models for anomaly classification within Microsoft's financial ledgers

Microsoft Explore Intern (SWE and PM) May 2021 – Aug. 2021
Microsoft *Redmond, WA*

- Used Azure Development Environment and Kusto Query Language to build anomaly detecting models for subscriptions
- Created reports for models in PowerBI and set up incident alerting upon anomaly detection in Jarvis
- Wrote code in C# (.NET Core) to analyze renewal failures by system error type

CS Department Teaching Assistant Aug. 2020 – May 2022
Wellesley College Computer Science Department *Wellesley, MA*

- TA'd and graded for CS232, the artificial intelligence course, and CS111, the introductory Python course

Projects and Research

Quantifying Privacy Harm via Predictive Privacy Submitted
First Author, Technical Paper (with Zhibin Shen and Steven M. Bellovin) *New York, NY*

- Introduced a new theoretical framework reconceptualizing privacy violations as predictive harms, where machine learning models infer sensitive traits that were never disclosed.
- Designed a mathematical harm model integrating probabilistic inference, contextual norms, and observer identity to formally describe how privacy injury evolves over time.
- Conducted large-scale empirical studies using synthetic population data and national surveys to quantify perceived harm across demographic and contextual scenarios.
- Built a supervised learning pipeline to estimate a non-analytic harm function, capturing the relationship between inference accuracy, attribute sensitivity, and social visibility.

Beyond Creepiness: Predictive Privacy In Progress
Lead Author, Law Review Article (with Steven M. Bellovin) *New York, NY*

- Analyzed how traditional privacy paradigms—based on explicit identifiers—fail to protect individuals against modern machine learning-based inferences, highlighting regulatory blind spots in existing US frameworks.
- Proposed the "Predictive Privacy" model to quantify and articulate privacy harm arising from *inferred* personal attributes, addressing novel types of injury not captured by current laws.
- Mapped concrete legal consequences for consumers and organizations by tracing how predictive harms (e.g., inferences about ethnicity, orientation) evade the protections of statutes like the U.S. Privacy Act.
- Outlined actionable recommendations for lawmakers and courts to modernize legal definitions of privacy harm, influencing ongoing policy debates.

Predictive Privacy: Master's Thesis May 2025
Columbia University *New York, NY*

- Designed and implemented the `Predictive Privacy` open-source library, introducing a framework to quantitatively assess privacy harms and support regulatory/legal claims of concrete injury.
- Engineered a synthetic, population-scale database using differential privacy methods to model sensitive attributes and simulate real-world data breaches.
- Developed and evaluated semi-supervised machine learning models that predict individual-level privacy risks under a variety of data exposure scenarios.

- Collaborated with legal experts to align technical definitions of privacy harm with evolving standards in U.S. privacy law and regulatory policy.
- Technical paper manuscript in progress. First author.

Comparing and Evaluating ChatGPT's Performance Giving Financial Advice 2024
Journal of Emerging Investigators *Fremont, CA*

- Analyzed ChatGPT's financial advice in comparison to real user Q&A threads on Reddit, focusing on accuracy, clarity, and overall quality of responses.
- Mentored high school students in conducting this study, published in the Journal of Emerging Investigators.
- Citation: Samant, S., Dhar, A., Kochar, S., Sreerama, A., Wang, A., & Sreerama, A. (2024). *Comparing and evaluating ChatGPT's performance giving financial advice with Reddit questions and answers*. Journal of Emerging Investigators. Retrieved from https://emerginginvestigators.org/articles/23-296.

Voice-Cloning AI: Understanding Legal Implications Using the Voiceprint Definition Dec. 2023
Columbia University *New York, NY*

- Created a comprehensive study on the use of AI in voice cloning, analyzing the capabilities of state-of-the-art technologies like Microsoft's VALL-E to synthesize personalized speech from minimal data input.
- Proposed a pioneering legal framework for assessing the implications of AI-generated voice cloning in cases of fraud and rights to publicity, outlining a procedure for evaluating damages

PACISCA: Probablistic Analysis of Confocally Imaged Synaptic Calcium Activity Oct. 2020
MIT Littleton Lab Research Project *Cambridge, MA*

- Co-authored a paper detailing ML techniques to detect synaptic regions of interest in Drosophila brain images.
- Paper accepted by MIT's IEEE URTC conference to be presented and published

Awards, Honors, and Achievements

- Appointed by Professor Norman Sadeh to join the *Smart City Privacy Technologies* project (funded by the National Science Foundation, 2025–2028).
- Law paper draft "Beyond Creepiness: Predictive Privacy" accepted to Privacy Law Scholars Conference (PLSC) 2025. Invited to present.
- Invited to roundtable on law and computer science. Hosted at University of Pennsylvania (2025).
- Accepted to attend The Cornell, Maryland, Max Planck Pre-doctoral Research School in Computer Science (CMMRS) 2025, with a EUR €1300 travel stipend.
- Received a $750 NSF travel grant to attend ACM Symposium (CS&Law '25)
- Invited to participate in Stanford HAI's invite-only "World Wide Knowledge AI Assistant" workshop in person
- Invited to attend Stanford HAI's invite-only "Trusting Digital Content in the Age of AI" conference in person
- Received a scholarship covering hotel, ticket, and expenses to attend TrustCon 2024
- Accepted to attend "Designing Safe(r) Digital Intimacy" Workshop @ Berkman Klein Center for Internet & Society. White paper in progress.
- Selected to supervise/advise an undergraduate student for research by Professor Steven M. Bellovin
- Invited to represent Professor Steven M. Bellovin at a roundtable on law and computer science. Hosted at University of Pennsylvania (2024).
- Selected as a Data Science Institute Scholar at Columbia. Received a $3000 research stipend.
- Received a $1000 NSF travel grant to attend ACM Symposium (CS&Law '24)
- Received a $500 travel grant to attend Stanford Treehacks by Stanford CS
- Passed Azure Certified exams for AI Engineer and Azure Fundamentals

Activities

Journal of Privacy and Confidentiality | *Reviewer* Oct. 2024 – Present
Journal of Emerging Investigators | *Associate Editor* Feb. 2024 – Present

Technical Skills

Skills: Java, Python, SQL, Kusto Query Language, PowerBI, .NET Core, Jarvis, R, Flask, C/C++, x86, JavaScript, HTML, CSS, MATLAB, Google Apps