I am fascinated by the intersections between natural language processing (NLP), human-computer interaction, and the social sciences and humanities. My primary research goal is developing socially aware language models to advance human-centered NLP, particularly:

- **NLP Normative Alignment:** How can we expose existing social biases and align language models to diverse human sociocultural norms?
- **NLP for Social Good:** How can human-centered NLP benefit end-users in social domains?

My academic training in computer science, math, cognitive science, and linguistics, along with my background in social identity philosophy, political science, and ethics makes me uniquely comfortable with navigating between these domains.

NLP Normative Alignment Efforts in aligning language models to human preferences, including reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), have proven capable. However, I believe robust normative alignment requires (1) systematic ways to evaluate encoded latent biases, and (2) stable techniques to incorporate complex human preferences into existing alignment pipelines.

(1) In pursuit of value-aligned NLP, I approached Professor Mark Riedl to join the Human-Centered AI (HCAI) Lab at Georgia Tech. While performing literature reviews, I became disturbed that we have ways to *evoke* biases in large language models (LLMs), but not systematic ways to *measure* them, nor *predict* when they occur. As a push towards systematic evaluation of LLM biases, for my senior thesis I developed a pipeline to prompt LLMs with automatically generated knowledge graphs (KGs) from movie scripts for character analysis, finding that LLMs do not have robust integration of sociocultural norms. I continued this work by helping fine-tune T5 models using precondition and effect information from LLM-generated KGs to act as world models in downstream RL tasks (*COLING 2025)*. Working on these projects, I grew fascinated with understanding what internal representations LLMs learn about the world, and especially the encoded latent human biases.

(2) While working with the HCAI lab, I noticed stories clustered *thematically* when embedded in 2D latent space, helping contrastively train a preference reward model for RLHF. I led human evaluation of the fine-tuned model, which outperformed models 20x larger for topic preferences and moral alignment (*arXiv 2022)*. However, the reward signals from our preference model were coarse; I was concerned that the sparse, scalar rewards in RLHF would limit robust alignment with complex human preferences. To address this, I am collaborating with Professors Mingyi Hong and Dongyeop Kang at The University of Minnesota to develop dense reward signals from multi-dimensional human preferences. I developed a pipeline to distribute interpretability values from LIME, SHAP, and integrated gradients at the token-level to shape an underlying reward model, helping reveal *which tokens* influence LLMs the most *for which* alignment task. These projects familiarized me with researching different RLHF formulations, and how nuances in reward models, or smart alterations to reward signals, can have incredible effects on fine-tuned models. I learned to conduct thorough human evaluation of LLMs, which I believe is merely a steppingstone towards better evaluation criteria.

NLP for Social Good With the current zeitgeist of LLM benchmarking, it is easy to proclaim a new model is more "aligned" by beating SOTA for arbitrary social knowledge metrics. End-users of language technologies are hardly a monolith. Thus, it is paramount to examine the effects of LLMs *in situ*.

To investigate social domains to *deploy* advances in NLP, I participated in the Research Experience for Undergraduates in Software Engineering REU at Carnegie Mellon University, where I worked with Dr. Norman Sadeh on usable privacy policies. I built an ensemble model of logistic regression classifiers and BERT-, GPT-, and Llama-based LLMs to predict collection of 38 different personal data types in privacy policy disclosures, indicating that many Android Apps may have *potential compliance issues* with disclosure requirements, ultimately harming end-users. These results led to a poster and second-authored paper (*Springer Computing 2024)*. I learned how architectural choices in NLP systems can directly affect users, and I strive to be constantly aware of *who* is affected by my research, in *what ways*.

At CMU, I have continued investigating human-facing situations in which LLMs underperform. While collaborating with Tom Norton, a law professor at Fordham University, we found that SOTA LLMs could correctly reason about omitted information in privacy policies under different regulatory requirements (GDPR, CCPA) *just 5.36%* of the time (*WISE 2024a)*. A human study also revealed that LLMs could provide accurate answers to everyday cybersecurity questions, but struggled with answers that were motivating, understandable by diverse groups, or actionable for users from varied backgrounds. However, careful prompt engineering could improve performance by *as much as 52.02%* (*WISE 2024b)*, which was further verified by an *in situ* study with 51 everyday users and 1050 questions (*USEC 2025 under review)*. These projects revealed deficiencies in LLMs that must be addressed before widescale adoption in more pressing social domains. I learned the importance of legal precision and collaborations with social scientists in NLP. Preparing these first-author publications and presenting these results at WISE 2024 have helped me develop the skills to communicate research with people from diverse backgrounds.

Future Goals Ultimately, I hope to be a professor, allowing me to teach, lead research projects, and advise students. One of my most fulfilling experiences in college was serving as a tutor and mentor for over 1000 students in the Online Master's in Analytics program at Georgia Tech for two years with Dr. Richard Vuduc. The tutoring program was designed for master's students who did not previously have access to formal computer science education. I hosted hourly sessions four times a week for students from across the world. My proudest moment was when an older student who felt stuck in his career asked me how to break into AI research because he was so deeply interested in the ML algorithms we were discussing. I hope to continue teaching both during my PhD and as a professor. I am confident that the University of Washington is the right institution for me to pursue my goals. At the University of Washington, I aim to apply my *people-first* approach to both theoretical and applied NLP.

I am interested in working with **Professor Aylin Caliskan** on the societal impacts of NLP. I am particularly interested in her work in *identity* and *representation* in language technologies. I believe I could leverage my background in critical race theory and identity philosophy to advance this research. It is still not entirely clear how to systematically identify and quantify biases in language models, and even more so how to develop sociocultural alignment techniques to steer language models away from these biases, and I am confident that Professor Caliskan's expertise in NLP and ethics would help me further investigate these questions. I would also be interested in working with **Professor Martin Saveski** to leverage NLP in developing algorithms that can encode societal values. A recently emerging field in NLP is pluralistic alignment, and I would be eager to investigate how to apply concepts from pluralism in AI systems, such as multi-dimensional/multi-objective RLHF and diverse human benchmarks, to social media algorithms under his guidance. Given UW's positioning as a leader in NLP, and especially as a pioneer in pluralistic alignment, along with its dedication to collaborations with other domains in information science, I believe that it is a strong fit for me to pursue my PhD.