Haoran Zheng   Web: haoranzhengraul.github.io   Email: raulzhenghaoran@gmail.com

Statement of Purpose

For as long as I can recall, I have been driven by an urge to understand how complex systems operate. This curiosity eventually led me to the field of Explainable AI, where exploring the intricacies of Machine Learning models became both a passion and a pursuit. This journey was highlighted in my recent paper, *SCENE: Evaluating Explainable AI Techniques Using Soft Counterfactuals*, in which I proposed a novel method to assess the performance of Explainable AI techniques in text classification. To further my interest in interpretability, I have explored a wide range of academic disciplines and held various positions as a teaching assistant (TA) and research assistant (RA), gaining valuable experience in creating course materials, teaching, and facilitating research processes. I aspire to continue studying advanced Machine Learning systems and Explainable AI as a university professor. Pursuing a Ph.D. in Computer Science at MIT would enable me to deepen my research passions, contribute meaningfully to the field, and work towards building a world of trustworthy AI systems.

Finding My Research Interests. Earning degrees in both STEM and social sciences has enabled me to approach research problems from multiple perspectives, ultimately helping me pinpoint my interest in Machine Learning and Explainable AI. With undergraduate degrees in Finance, Economics, and History, and two years of full-time experience working with large heterogeneous data at China Automotive Systems, Inc., I have a solid foundation in both theoretical and applied research. Most recently, I graduated from the M.S. in Applied Data Science program at the University of Chicago in December 2024 with a 4.0 GPA. For my master's thesis, advised by Professor Utku Pamuksuz, I conducted a comprehensive empirical evaluation of the ten most cited Explainable AI methods, such as LIME and SHAP, across three popular neural network architectures—Transformers, CNNs, and RNNs—in text classification. The final results were presented to faculty members at the program's capstone presentation.

While researching evaluation metrics for Explainable AI methods, I recognized limitations in existing approaches. Motivated by filling these gaps, I proposed the SCENE method after successfully defending my thesis. Within a month, I completed the first draft of the paper on my own. SCENE addresses two significant challenges in Explainable AI. First, traditional counterfactual-based methods rely on flipping the prediction label, which is not always feasible. To overcome this, SCENE introduces the concept of 'soft counterfactuals,' where the outcome is measured by changes in model output probabilities, while also accounting for input distances. Second, existing erasure-based methods, like comprehensiveness, and perturbation methods, like infidelity, often face plausibility issues because the altered inputs can fall outside the original data distribution. SCENE tackles this by leveraging state-of-the-art Large Language Models (LLMs) for token-level substitutions, ensuring that the modified inputs are contextually appropriate and semantically meaningful.

During those 17-hour workdays over academic breaks spent writing the SCENE paper—when I set my own deadlines—I often debated ideas even in my sleep, embracing the independence to fully explore my concepts. Yet, I remained happy and motivated, proactively seeking opportunities to discuss my ideas with peers. This intense period solidified my passion for Explainable AI and the research process itself. Recognizing the fulfillment I found in this work, I know that pursuing a Ph.D. in Computer Science is the next step for me.

The initial version of my paper attracted interest from the field. The Causal Inference Workshop at KDD 2024 approached my advisor to invite me as a last-minute substitute speaker at their workshop in Barcelona, Spain. Although the opportunity excited me, it did not

materialize due to scheduling conflicts. Following this, I submitted the paper to conferences with archival tracks to receive rigorous peer-reviewed feedback. The first submissions included KDD 2025, the EMNLP 2024 Workshop BlackBoxNLP, and the NeurIPS 2024 Workshop InterpretableAI. As of writing this statement, only BlackBoxNLP has provided feedback. The reviewers praised the topic's timeliness, the comprehensive literature review, and the depth of the experimental design. They also pointed out areas for improvement, such as the article's structure and overly complicated illustrations, as well as some finer aspects of the experimental design. I am currently revising the paper to address these concerns and aim to publish the improved version before starting the Ph.D. program in 2025.

In addition to my main research, I strive to understand the fundamental challenges of black-box systems across various disciplines. To this end, I currently hold two RA roles in the field of Behavioral Science, where the ultimate black-box system is the human mind. At the Mindworks Lab within the Roman Family Center for Decision Research (RF-CDR) at the UChicago's Booth School of Business, I co-manage studies and collect data for renowned scholars like Dr. Nicholas Epley and Dr. Sarah Sebo. Notably, in Dr. Sebo's study, *Engaging with a Robot to Improve Your Overall Wellbeing*, I collaborated with the UChicago Human-Robot Interaction (HRI) Lab. We investigated whether individuals whose personalities closely match those of the robots they interact with benefit more from positive psychology exercises and find the interaction more enjoyable. My other RA role is at the Epley Lab, also within RF-CDR, where I work directly with Dr. Epley and coordinate the Multicultural Study alongside Yale scholar Yin Li, investigating how people from different cultural backgrounds gauge others' beliefs and preferences. Additionally, I serve as a History RA at Ohio State University with my former mentor, Dr. Christopher Reed, applying modern NLP techniques and the SCENE method to address traditional historical research questions, such as identifying the significance of previously unseen documents.

Equal Access in STEM. I am genuinely excited to resume my teaching journey as a doctoral student—a passion that began during my undergraduate studies at OSU. During this time, I worked as a TA for Statistics 2320 under Professor Bonnie Schroeder. This sophomore-level course, which is a requirement for over 900 students each semester from diverse backgrounds, required me to lead two weekly recitations of 65 students each, covering topics ranging from hypothesis testing to ANOVA. Over four semesters, I honed my skills in conveying complex material to students with varying proficiency levels, using R in active problem-solving sessions, and supporting students with customized review notes and visual aids. I firmly believe in breaking down barriers to knowledge so that all students feel empowered to pursue careers in STEM. In recognition of my dedication, OSU honored me with the Pace Setters Award, the Fisher College of Business's highest distinction for excellence in academics, leadership, and service.

Beyond academia, my industry experience at China Automotive Systems, Inc. reinforced my commitment to practical problem-solving. I designed a machine learning model that combined large sets of tabular and textual data to forecast production and identify potential issues. This hands-on experience taught me the importance of adaptability and creativity—qualities I aim to bring to my research and interactions within MIT's diverse academic community. Additionally, my professional experience have provided me with a stable financial foundation to support my PhD journey. I am able to potentially self-fund my PhD for up to two years. I also plan to apply for several fellowships, including the Open Phil AI Fellowship, to further ensure that I can fully dedicate myself to my research and academic development.

Outside of my professional life, I enjoy connecting with people from various backgrounds through my passion for sports. I have played for the Columbus Crew's reserve soccer team, competed as a kickboxer, and even contested as a professional e-sports player. These experiences have enabled me to break the ice with students and colleagues, fostering an environment where everyone feels included and valued.

My personal journey has instilled in me the courage to persevere through any challenges and defy seemingly impossible odds. In 2017, I was diagnosed with severe depression, facing a year of unbalanced diet, irregular sleep, and suicidal thoughts that made daily functioning a constant struggle. Discussing this period isn't easy, even now, but with the unwavering support of family and friends, I began the slow process of healing. Seeking professional help, I returned to my studies with renewed determination—not only to catch up but to excel. Over nine semesters, I completed three majors while maintaining a 3.8 GPA. This experience marked a turning point, teaching me invaluable lessons about resilience, the importance of mental health, and the power of a supportive community. Overcoming this adversity didn't just restore me to where I was—it propelled me forward, shaping me into a stronger, more compassionate individual ready to face future challenges head-on. It also deepened my empathy for others facing similar struggles. Now, as an advocate for mental health awareness, I strive to create supportive environments wherever I go. These qualities will empower me to contribute positively to MIT's community and inspire others to confront their challenges with the same confidence and resilience.

Where I See Myself. The above experiences underscore my motivation to join MIT's Ph.D. program in Computer Science. I aspire to make meaningful contributions to Explainable AI, and MIT offers unparalleled resources. Instead of submitting a paper to a conference and waiting months in the dark for feedback, I would have the opportunity to receive regular guidance from renowned scholars like Dr. Jacob Andreas and Dr. Yoon Kim. My recent paper focused on the automatic interpretation and explanation of learned models, as well as human-like language understanding, which closely aligns with Dr. Andreas's research interests. It would also be an honor to work with Dr. Yoon Kim, as his expertise in understanding the capabilities and limitations of language models would provide me with valuable insights into the complex NLP challenges I am interested in. I recently read *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*, a fascinating discussion on causal inference and language models co-advised by both professors, which strongly resonated with my research. I am confident that I can contribute to the missions of both research groups and help expand the scope of research in Explainable AI.