# Haotian Chen

*Research Statement*

## Research Experience

My research aims to address the question, "Why did existing AI models fail to learn the decision rules of humans or to think like humans?" My past research has helped to answer this question by formulating the problem, diagnosing the trustworthiness of language models, and exploring the underlying causes. Specifically, through many failures, I kept optimizing my research paradigm until the current version: 1) investigating deep models' decision rules, 2) identifying the underlying patterns of the rules, and 3) analyzing the underlying causes before 4) improving models. I identify AI's problems (e.g., discrimination, neglecting rationales, poor generalization ability, etc.) and their causes (e.g., data imbalance, lack of causal information, incomplete testing data, etc.), which guides me to propose effective solutions based on statistics, causality, and neural language processing techniques, improving models' abilities in some downstream tasks (e.g., relation extraction, legal judgment prediction, text summarization, etc.).

• **Motivation for Postdoc**: Since my life goal is to lead a team (e.g., a lab) to change the world through ML models, I'm strongly motivated to become a qualified leader with a broad vision and deep insights. As I have always been self-directing my own research in CS up to now, my primary concern is that I've no idea about how to conduct world-class research in CS, and I'm also not confident about my research paradigm. I have never had a deep collaboration with world-class researchers in CS to deepen my insights, broaden my vision, or perform "imitation learning". I've realized that conducting research in isolation is inefficient. Often, teammates are like walking knowledge repositories, and the accumulation of information and depth of thought can be greatly expanded through communication. As a result, everyone is learning and growing together, making the team stronger. This, in turn, gives us the opportunity and ability to conduct world-class research.

## Research Goals

I believe that conducting research on the intersection between machine learning and its real-world applications (e.g., NLP, CV, Fintech, etc.) will satisfy my interest the most: understanding more about ML tools makes me able to contribute to a wide range of tasks I'm interested in. Therefore, I will strive to work on the following problems with large societal impacts in the next few years.

• **Causality Learning Ability**: How can we build models that make decisions according to rationales (causality)? What is the right way (e.g., feature attribution) to characterize the decision rules of models? How can we recognize (evaluate) those causality-based models that can be deployed in real-world applications with strong robustness, generalization ability, trustworthiness, and performance?

• **Data Imbalance**: If each training sample contains only causal information and the corresponding label, models will easily learn to make predictions according to causality. How can we distinguish causal and non-causal information? How to eliminate non-causal information as much as possible?

• **Model Architecture**: As for a simple linear model, it can easily learn the accurate causal effect from noisy training data if we set the coefficient of its irrelevant variables to zero. How can we design the architecture of a learning model to make it able to distinguish causality from spurious correlations? How can human knowledge be infused into learning models?

• **Reward Specification (Objective Function)**: Human values are too complex to be specified by hand. How can we infer complex value functions from data? Are the functions able to guide models to learn causality? How should an agent make decisions when its value function is approximate due to noise in the data or inadequacies in the model?

• **Optimization**: Sampling distorts the original information, rendering the sampled data imbalanced. Since ML methods greedily absorb all the correlations in the training data to minimize their training error, the randomness of optimization processes (e.g., stochastic gradient descent) renders "model learns causality" becomes a random event. How can we maximize the probability of the event by controlling the random optimization process? Can we thereby find out the fragile parts of models?

• **A Recent Idea about Conducting Research in LLM Agent through Information Theory**: The core of improving an Large Language Model (LLM) Agent's performance lies in the prompt, as the quality of the prompt significantly determines the subsequent results. In other words, it's about what information you initially input. The information you input reduces the entropy, allowing the LLM to know which sub-space within its whole world knowledge space to search for answers, and then provide feedback. In my opinion, the nature of chain-of-thought approach is involving two LLMs interacting step by step to further reduce the LLM's entropy. When the entropy is reduced, the results of the LLM tend to become more certain and stable. Stability means a smaller variance while providing the results people want, thus improving the performance. So, can we provide a conceptual framework (with theoretical analysis) to explain such an assumption and give experimental demonstration accordingly? I consider it as an exciting work. Perhaps the conceptual framework can guide the future of prompt engineering. For example, given a specific task and requirement, what conditions in a mathematical sense does a natural language prompt need to meet in order to elicit a stable and correct response from a LLM Agent?

## Research Plan

I strive to keep learning and growing to tackle any of the aforementioned problems. I'm eager to be involved in more discussions and brainstorms to update my understanding or even reconstruct my research goals. Currently, I try to discuss these problems in specific applications such as relation extraction (RE) and legal judgment prediction (LJP). I construct two benchmarks where the decision rules of humans are annotated in document-level RE (DocRE), and the out-of-distribution (OOD) samples are found and collected to evaluate DocRE methods. The benchmarks reveal the bottleneck of existing models. By further applying feature attribution methods to obtain the decision rules of ML models, I have many important findings through extensive experiments. These findings contribute to my research goals and also drive me to rethink the development of ML models and thus explore many potential solutions. The details are presented in the papers.

• **Future Work**: In the future, I have to rethink many research problems in the background of very large models, including but not limited to developing methods to figure out the decision rules of very large models, exploring the performance boundary of large models in some attractive real-world applications, designing a comprehensive benchmark to evaluate various kinds of ability of very large models (e.g., robustness, generalization ability, causality learning ability, performance, etc.), developing methods to improve all kinds of ability of very large models.