As large language models (LLMs) become more powerful and are increasingly deployed in daily life, social contexts, and even high-stakes scenarios, the **risks** involved are also becoming more significant. This has shaped my research interests toward **LLM safety** - both the **inherent safety of LLMs** and their safety **when deployed in real-world complex scenarios, interacting with real humans and society (e.g., as LLM agents or LLM systems)**, with a focus on:

i) **Ensuring the safety of LLMs:**

i.i) Non-Adversarial Safety - How can we prevent LLM failures in edge/accidental cases or due to inherent vulnerabilities, e.g. issues with robustness, fairness, privacy, interpretability, hallucination, and overall trustworthiness?

i.ii) Adversarial Safety - How can we protect LLM systems from malicious use and activities, e.g. jailbreaking, the production and propagation of misinformation, training data extraction, etc.?

ii) **Ensuring the safety of LLMs in complex scenarios (i.e., as LLM agents/systems or in social simulations):**

ii.i) Safety of LLM Agent & System - How can we prevent LLMs from taking harmful actions when deployed as agents or systems in real-world scenarios and social contexts, with potentially irreversible / catastrophic consequences?

ii.ii) Safety of Social Simulation - As LLMs and (multi-)LLM agents are increasingly used to simulate social interactions and large-scale human behavior, how can we ensure the safety and reliability of these simulations?

I Ensuring the Safety of LLMs

In recent years, LLMs are becoming increasingly powerful and prevalent. However, as their capabilities grow, the potential consequences of failure also become more severe, especially considering their deployment in high-stakes scenarios. This highlights the importance of ensuring LLM safety, both in terms of mitigating risks within the LLM itself and defending against malicious external attacks.

I.I Non-Adversarial Safety

Solving the issue of LLM safety at its core requires first addressing the inherent risks within the models themselves, including limited robustness to long-tail behavior and out-of-distribution (OoD) inputs, unfairness toward underrepresented populations, the lack of trustworthiness in LLM predictions due to their black-box nature (interpretability and explainability), as well as the risks of privacy leakage and hallucination that could lead to misinformation, etc. Below, I outline some of my previous efforts to tackle these challenges:

Dialectal Robustness via Feature Composition Existing LLMs that mainly focus on standard languages like Standard American English (SAE), which often demonstrate significantly reduced performance when applied to OoD inputs from non-standard dialects. In my EMNLP 2023 papers [4][5], I incorporated linguistic knowledge to address this distribution shift, e.g., introducing Dialect Adaptation via Dynamic Aggregation (DADA) [4], a modular approach that adapts SAE models to multiple dialects simultaneously from a finer-grained perspective — non-standard linguistic features — by enhancing dialectal robustness through the composition of specific linguistic feature adapters, while also providing strong interpretability. Additionally, in earlier work, I made a similar attempt to improve model robustness by mitigating multiple spurious correlations in datasets through parameter-efficient adaptation & composition [6].

Fairness for Underrepresented Subgroups While many previous works have employed LLMs for classification tasks on tabular data, fairness-related issues in this context remain relatively underexplored, especially considering the extensive use of tabular data in high-stakes domains. In my NAACL 2024 work [3], I addressed this gap by conducting a series of experiments to investigate the fairness of using LLMs for tabular predictions, comparing them with traditional ML models. Our findings demonstrate that LLMs tend to inherit biases from their pretraining data, with significant fairness gaps persisting for minority subgroups, even with common bias mitigation methods. While label-flipping of in-context examples can help "unlearn" these biases, it comes at the cost of significantly degrading the model's accuracy.

I.II Adversarial Safety

The situation becomes even more concerning when access to LLMs is made available to the public, potentially exposing them to malicious actors. This leads me to explore how we can protect LLM systems from the malicious activities, including jailbreaking, the production and propagation of misinformation, training data extraction, etc.

Understanding Jailbreaking with Component Attribution While LLMs have been trained with safety alignment to prevent harmful and undesirable generation, these safety mechanisms have proven to be bypassed with little effort through various jailbreaking attempts. Instead of proposing new attack or defense methods, which often turn into a cat-and-mouse game, I am more interested in analyzing the factors contributing to the success of such jailbreaking. In my

ongoing work, I seek to understand jailbreaking attacks from an interpretability perspective. By attributing jailbreaking to the most influential model components, we explored several interesting questions, e.g. whether jailbreaking attacks target specific subsets of model components or if they are more widespread, whether different attack types target different component subsets, and if so, whether there are transferable correlations. Additionally, I am interested in analyzing the dynamics of jailbreaking, such as when LLMs begin to become sensitive to these attacks during the training process.

Preventing Misinformation through Susceptibility Modeling As the cost of accessing and deploying LLMs becomes more affordable, it raises potential safety risks of large-scale misinformation production and propagation by malicious actors on social media, especially due to their strong persuasive capabilities, making it easier to manipulate opinions, spread hate, and incite violence. To address this challenge, in my EMNLP 2024 papers [2][8], I propose a computational approach to model users' latent susceptibility to misinformation, based solely on their posting and sharing activities on social media. Beyond investigating the underlying mechanisms of misinformation propagation on a large scale, our computational modeling enables the identification of users at high risk of believing false claims, allowing for proactive preventive measures to safeguard the online environment, fostering positive social impacts.

Beyond these, to ensure safer generation, I am excited to further explore unlearning techniques to fundamentally remove harmful knowledge embedded in LLMs, and investigate more robust safety alignment methods. Also, I am interested in using unlearning and retrieval-augmented generation (RAG) to enhance privacy & copyright protection.

II Ensuring the Safety of LLMs in Complex Scenarios

As LLMs are increasingly grounded and deployed as LLM agents, the associated safety risks are thus more pronounced and amplified, due to their often irreversible actions with potentially catastrophic consequences. In addition to the general attacks and vulnerabilities inherited from LLMs, LLM agents also introduce additional risks specific to their use.

II.I Safety of LLM Agent & System

The growing use of LLM agents in daily-life tasks and social contexts, such as web browsing, shopping, coding, scheduling meetings, negotiation, etc., has raised safety concerns among real-world users, especially regarding privacy.

Evaluating LLM Agent Privacy Awareness In my co-authored paper [1], we proposed a novel framework, along with extensible datasets, to benchmark LLMs' privacy awareness when deployed as LLM agents for assisting users with everyday tasks. With this framework, we demonstrate a discrepancy of privacy norm awareness between LLMs' performance in answering probing questions and their actual behavior when executing user instructions in an agent setup.

Furthermore, as (multi-)LLMs are increasingly integrated into complex systems (i.e., as LLM systems), the capabilities of such systems increase explosively with the capabilities of each individual LLM; however, same are the corresponding safety risks (i.e. combinational safety). In addition to designing better coordination/collaboration mechanisms for LLM systems, the complexity of such systems makes the evaluation and assurance of safety and reliability more challenging than ever before. This has sparked my profound interest, particularly in exploring the capabilities of LLM agents and LLM systems for solving complex tasks and, more importantly, in evaluating and ensuring their safety with sandbox environments, given the potential risks associated with the irreversible consequences of their actions.

II.II Safety of Social Simulation

Additionally, as LLMs and (multi-)LLM agents are increasingly used to simulate social interactions and even large-scale human behavior [2][8], the associated safety risks and reliability concerns have become an emerging problem. Before employing these social simulations for the study of various social phenomena or for scientific discoveries in social science research, how can we make sure that the robustness, fairness, and even interpretability of these simulations are thoroughly evaluated, and that their ability to make reliable decisions in complex social contexts is rigorously assessed?

III Why?

Looking ahead, my long-term goal is to leverage my experience in NLP, ML, and CSS to contribute to the development and deployment of **safer and more trustworthy** AI systems, with a careful consideration of their impact on **humans and society** in **real-world scenarios** - a challenge that becomes increasingly critical as LLMs grow more powerful and deeply integrated into complex systems.

IV References

[1] Yijia Shao, Tianshi Li, Weiyan Shi, **Yanchen Liu**, and Diyi Yang. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action.
In Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

[2] **Yanchen Liu**, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, Diyi Yang. Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach.
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)

[3] **Yanchen Liu**, Srishti Gautam, Jiaqi Ma, Himabindu Lakkaraju. Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classifications.
In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)

[4] **Yanchen Liu**, William Held, Diyi Yang. DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules.
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

[5] Zedian Xiao, William Held, **Yanchen Liu**, Diyi Yang. Task-Agnostic Low-Rank Adapters for Unseen English Dialects.
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

[6] **Yanchen Liu***, Jing Yan*, Yan Chen*, Jing Liu, Hua Wu. SMoA: Sparse Mixture of Adapters to Mitigate Multiple Dataset Biases.
In ACL 2023 Workshop on Trustworthy Natural Language Processing (ACLW 2023)

[7] **Yanchen Liu**, Timo Schick, Hinrich Schütze. Semantic-Oriented Unlabeled Priming for Large-Scale Language Models.
In ACL 2023 Workshop on Simple & Efficient Natural Language Processing (ACLW 2023)

[8] Mingyu Derek Ma, Alexander K. Taylor, Nuan Wen, **Yanchen Liu**, Po-Nien Kung, Wenna Qin, Shicheng Wen, Azure Zhou, Diyi Yang, Xuezhe Ma, Nanyun Peng, Wei Wang. MIDDAG: Where Does Our News Go? Investigating Information Diffusion via Community-Level Information Pathways.
In Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024 Demonstrations)

[9] Qi Wu, Chong Zhang, **Yanchen Liu**. Custom Sine Waves Are Enough for Imitation Learning of Bipedal Gaits with Different Styles.
In Proceedings of the 2022 IEEE International Conference on Mechatronics and Automation (ICMA 2022)