# Certifiably Robust Learning via Knowledge-Enabled Logical Reasoning

**Bo LI**

**Department of Computer Science, University of Illinois at Urbana–Champaign**

**Email: lbo@illinois.edu**

The ubiquity of intelligent systems underscores the paramount importance of ensuring their trustworthiness. Traditional machine learning approaches often assume that training and test data follow similar distributions, neglecting the possibility of adversaries manipulating either distribution or natural distribution shifts, which can lead to severe trustworthiness issues in machine learning. The speaker's previous research has demonstrated that motivated adversaries can circumvent anomaly detection or other machine learning models at test-time through evasion attacks, or inject malicious instances into training data to induce errors through poisoning attacks. In this talk, the speaker will provide a succinct overview of our research on trustworthy machine learning, including robustness, privacy, generalization, and their underlying interconnections. The speaker will showcase the trustworthiness issues of large language models, including GPT-4 and GPT-3.5. She will then discuss the current state of the art in certifiably robust defenses based on purely data-driven models and demonstrate that they have reached a bottleneck. The speaker will present her team's recent research on *certifiably robust learning via knowledge-enabled logical reasoning*, showing that it is possible to: 1) certify the robustness of such an end-to-end framework and significantly improve the certified robustness on large-scale datasets, 2) prove that such a framework is more robust than a single data-driven model under mild conditions, and 3) scale it for a variety of downstream tasks such as image classification, information extraction, PDF malware classification, and data generation.