# Certified Trustworthiness in the Era of Large Language Models

**Linyi Li**

Simon Fraser University
School of Computing Science
linyi_li@sfu.ca

## Abstract

Along with the broad deployment of deep learning (DL) systems, their lack of trustworthiness, such as their lack of robustness, fairness, and numerical reliability, is raising serious social concerns, especially in safety-critical scenarios such as autonomous driving and aircraft navigation. Hence, a rigorous and accurate evaluation of the trustworthiness of DL systems is essential and would be a prerequisite for improving DL trustworthiness. **The first part** of the talk will be an overview of certified methods for DL trustworthiness. These methods provide computable guarantees for DL systems in terms of worst-case trustworthiness under certain realistic conditions, such as the accuracy lower bound against arbitrary tiny perturbations. Based on our taxonomy and systematization, we illustrate key methodologies, specifically semantic randomized smoothing and branch-and-bound, and their implications for certified DL trustworthiness.

As a representative of recent DL breakthroughs, large language models (LLMs) are transforming our lives, but, on the other hand, posing more challenges to trustworthiness. For example, LLMs can be jailbroken with adversarial prompts to output harmful content with bias, harassment, misinformation, and more. **The second part** of the talk will be an overview of LLM trustworthiness. We will start with sharing hands-on experience in developing fontier LLMs, then illustrate common LLM trustworthiness issues via examples, then demonstrate evaluation challenges, take one benchmark as an example, and conclude by envisioning certifiable trustworthiness for LLMs.

The first part of the talk will be an overview of certified methods for DL trustworthiness following the structure of our SoK paper (Li, Xie, and Li 2023). We will introduce the taxonomy of certified methods, the key design methodologies, and the frontier in terms of state-of-the-art trustworthiness guarantees on common DL tasks. We will also dive deep into two representative methods: semantic randomized smoothing (Li et al. 2021) and branch-and-bound with differentiable linear relaxations (Zhang et al. 2022).

The second part of the talk will be an overview of LLM trustworthiness issues and the envisioning of certified methods for LLMs. We present some hands-on experience gained from developing industry-level frontier LLMs. Then, we illustrate common trustworthiness issues such as those in DecodingTrust (Wang et al. 2023) and more recent ones such as fine-tuning attacks (Qi et al. 2024). After that, we will highlight challenges and insights from building LLM benchmarks, taking InfiBench (Li et al. 2024) and HarmBench (Mazeika et al. 2024) as examples. We conclude by discussing how to scale certifiable methods for LLMs and presenting some preliminary results.



## References

Li, L.; Geng, S.; Li, Z.; He, Y.; Yu, H.; Hua, Z.; Ning, G.; Wang, S.; Xie, T.; and Yang, H. 2024. InfiBench: Systematically Evaluating the Question-Answering Capabilities of Code Large Language Models. In *Neural Information Processing Systems Datasets and Benchmarks Track*.

Li, L.; Weber, M.; Xu, X.; Rimanic, L.; Kailkhura, B.; Xie, T.; Zhang, C.; and Li, B. 2021. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 535–557.

Li, L.; Xie, T.; and Li, B. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*, 1289–1310. IEEE.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *Forty-first International Conference on Machine Learning*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, H.; Wang, S.; Xu, K.; Li, L.; Li, B.; Jana, S.; Hsieh, C.-J.; and Kolter, J. Z. 2022. General cutting planes for bound-propagation-based neural network verification. *Advances in neural information processing systems*, 35: 1656–1670.