# Research statement

**Mintong Kang**
mintong2@illinois.edu

## 1 Motivation

Deep neural networks (DNNs) have demonstrated impressive achievements in different domains. However, as DNNs are increasingly employed in real-world applications, concerns regarding their trustworthiness and reliability have emerged, including vulnerability to adversaries [9, 14] and bias/fairness issues [7, 1]. To address these concerns, it is crucial to provide the *worst-case certification* for the model predictions (e.g., certified robustness [3, 6]), and develop techniques that enable users to statistically evaluate the *uncertainty* linked to the model predictions (e.g., conformal prediction [13, 11, 4, 8]).

In particular, considering potential adversarial manipulations during test time, where a small perturbation could mislead the models to make incorrect predictions [12, 9], different robustness certification approaches have been explored to provide the worst-case prediction guarantees for neural networks. On the other hand, conformal prediction has been studied as an effective tool for generating prediction sets that reflect the prediction uncertainty of a given black-box model, assuming that the data is exchangeable [13, 11]. Such prediction *uncertainty* has provided a probabilistic guarantee in addition to the *worst-case certification*. However, it is unclear how such *uncertainty* certification would perform in the worst-case adversarial environments.

Although the worst-case certification for conformal prediction is promising, it is challenging for purely data-driven models to achieve high **certified prediction coverage (i.e., worst-case coverage with bounded input perturbations)**. Recent studies on the learning-reasoning framework, which integrates extrinsic domain knowledge and reasoning capabilities into data-driven models, have demonstrated great success in improving the model worst-case certifications [15, 16]. In this paper, we aim to bridge the **worst-case robustness certification** and **uncertainty certification**, and explore whether such knowledge-enabled logical reasoning could help improve the certified prediction coverage for conformal prediction. We ask: *How to efficiently integrate knowledge and reasoning capabilities into DNNs for conformal prediction? Can we prove that such knowledge and logical reasoning enabled framework would indeed achieve higher certified prediction coverage and accuracy than that of a single DNN?*

## 2 Background

**Randomized smoothing** [3] serves as a sound and efficient model robust certification framework. It first constructs a smoothed classifier by averaging the model predictions in the local Gaussian region, and then provides a certified radius which guarantees a consistent prediction in the local region with Neyman-Pearson lemma.

**Conformal prediction** [13, 11] is a statistical tool to construct the prediction set with guaranteed prediction coverage. Suppose that we have $n$ data samples $\{(X_i, Y_i)\}_{i=1}^n$ with features $X_i \in \mathbb{R}^d$ and labels $Y_i \in \mathcal{Y} := \{1, 2, ..., N_c\}$. Assume that the data samples are drawn exchangeably from some unknown distribution $P_{XY}$. Given a desired coverage $1 - \alpha \in (0, 1)$, conformal prediction methods construct a prediction set $\hat{C}_{n,\alpha} \subseteq \mathcal{Y}$ for a new data sample $(X_{n+1}, Y_{n+1}) \sim P_{XY}$ with the guarantee of *marginal prediction coverage*: $\mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})] \geq 1 - \alpha$.

**Probabilistic circuits** (PCs) [2] define a joint distribution over a set of random variables. PCs encapsulate a broad set of tractable probabilistic models and enable efficient computations of marginal probabilities. PCs achieve an impressive tradeoff between model expressiveness and inference efficiency, and thus, we leverage them to encode domain knowledge rules.

## 3 Method

In [5], we provide a certifiably robust learning-reasoning conformal prediction framework (COLEP) via knowledge-enabled logical reasoning. Our contributions span theoretical and empirical aspects.

- We propose COLEP, a data-driven learning with knowledge-enabled logical reasoning conformal prediction framework that enables reasoning inference via probabilistic circuits.
- We provide certification for probabilistic circuit structures and the certified prediction coverage for the end-to-end COLEP. We also provide certified coverage for COLEP considering the finite samples of the calibration set in conformal prediction.
- We theoretically prove that COLEP achieves higher prediction coverage and accuracy than that of a single model as long as the utilities of knowledge models are non-trivial.
- We empirically show that COLEP demonstrates higher certified coverage and prediction accuracy compared with SOTA baselines on GTSRB, CIFAR-10, and AwA2 datasets.

The COLEP is comprised of a data-driven *learning component* and a logic-driven *reasoning component*. The learning component is equipped with a main model to perform the main task of classification and $L$ knowledge models, each learns different concepts from data. Following the learning component is the reasoning component, which consists of $R$ subcomponents (e.g., PCs) responsible for encoding diverse domain knowledge and logical reasoning by characterizing the logical relationships among the learning models. The COLEP framework is depicted in Fig 1.

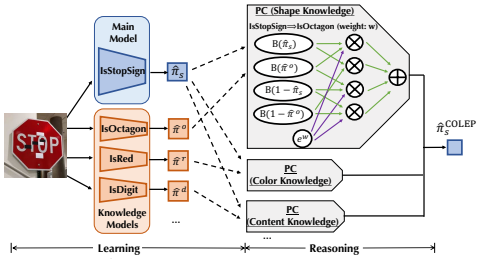
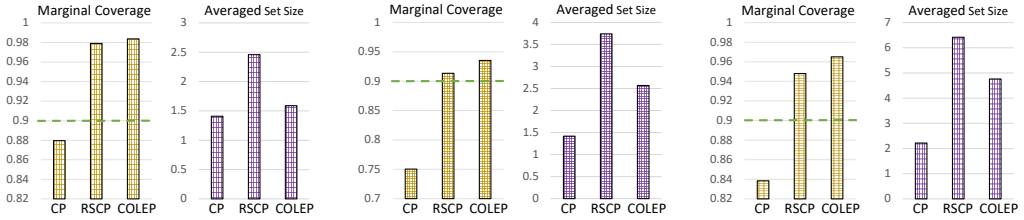



Figure 1: Overview of COLEP.

We conduct a thorough analysis for our COLEP framework. In Theorem 1 of [5], we certify the bounds for class probabilities after the correction of the reasoning component. Specifically, we apply randomized smoothing for bounds computation for the learning component, and explicitly propagate the bounds through the PCs. In Theorem 2 of [5], we provide a formulation of how to construct the conformal prediction set and certify its validity for the robustness guarantee. In Theorem 4 of [5], we prove that the COLEP demonstrates a higher prediction coverage than a single DNN as long as the utility of the knowledge models and PCs is non-trivial. In Theorem 5 of [5], we prove that COLEP demonstrates a higher prediction accuracy than a single DNN under mild conditions.

## 4 Results




(a) GTSRB (b) CIFAR-10 (c) AwA2

Figure 2: Comparison of the marginal coverage and averaged set size for CP, RSCP, and COLEP under PGD attack ($\delta = 0.25$) on GTSRB, CIFAR-10, and AwA2. The nominal coverage level (green line) is 0.9.

**Prediction Coverage and Prediction Set Size under Adversarial Attacks.** We evaluate the marginal coverage and averaged set size of COLEP and baselines under adversarial attacks. We compare the results with the standard conformal prediction (CP) and the SOTA conformal prediction with randomized smoothing (RSCP). We apply PGD attack [10] with the same parameters on CP, RSCP, and COLEP. The comparison of the marginal coverage and averaged set size for CP, RSCP, and COLEP under PGD attack ($\delta = 0.25$) is provided in Figure 2. The results indicate that the marginal coverage of CP is below the nominal coverage level 0.9 under PGD attacks as data exchangeability is violated, while COLEP still achieves higher marginal coverage than the nominal level, validating the robustness of COLEP for conformal prediction in the adversary setting. Compared with RSCP, COLEP achieves both larger marginal coverage and smaller set size, demonstrating that COLEP maintains the guaranteed coverage with less inflation of the prediction set. The observation validates our theoretical analysis that COLEP can achieve better coverage than a single model with the power of kwowledge-enabled logical reasoning.

## References


[1] Bhaskar Ray Chaudhury, Linyi Li, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. *NeurIPS*, 2022.

[2] Y Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. *UCLA. URL: http://starai. cs. ucla. edu/papers/ProbCirc20. pdf*, 2020.

[3] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[4] Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-rag: Certified generation risks for retrieval-augmented language models. *arXiv preprint arXiv:2402.03181*, 2024.

[5] Mintong Kang, Nezihe Merve Gürel, Linyi Li, and Bo Li. Colep: Certifiably robust learning-reasoning conformal prediction via probablistic circuits. *ICLR*, 2024.

[6] Mintong Kang, Linyi Li, and Bo Li. Fashapley: Fast and approximated shapley based model pruning towards certifiably robust dnns. *SaTML*, 2022.

[7] Mintong Kang, Linyi Li, Maurice Weber, Yang Liu, Ce Zhang, and Bo Li. Certifying some distributional fairness with subpopulation decomposition. *NeurIPS*, 2022.

[8] Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. Certifiably byzantine-robust federated conformal prediction. 2023.

[9] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *NeurIPS*, 2023.

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[11] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020.

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[13] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[14] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Liu, Yu Cheng, Sanmi Keyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *NeurIPS*, 2023.

[15] Zhuolin Yang, Zhikuan Zhao, Boxin Wang, Jiawei Zhang, Linyi Li, Hengzhi Pei, Bojan Karlaš, Ji Liu, Heng Guo, Ce Zhang, and Bo Li. Improving certified robustness via statistical learning with logical reasoning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

[16] Jiawei Zhang, Linyi Li, Ce Zhang, and Bo Li. CARE: Certifiably robust learning with reasoning via variational inference. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

## 5 Short list of references

The tutorial covers the following major references covering robustness certification and conformal prediction.

[1] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing, ICML 2019.

[2] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. NeurIPS 2020.

## 6 Long list of references

The references listed below are crucial to the advancement of my research.

[1] Weng, Lily, et al. "Towards fast computation of certified robustness for relu networks." International Conference on Machine Learning. ICML 2018.

[2] Zhang, Huan, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. "Efficient neural network robustness certification with general activation functions." NeurIPS 2018.

[3] Shafer, Glenn, and Vladimir Vovk. "A Tutorial on Conformal Prediction." JMLR 2008.

[4] Athalye, Anish, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." ICML 2018.

[5] Croce, Francesco, and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." ICML 2020.

[6] Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." CVPR 2018.

[7] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." ICLR 2018.

[8] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." ICLR 2021.

[9] Song, Yang, and Stefano Ermon. "Improved techniques for training score-based generative models." NeurIPS 2020.

[10] Nie, Weili, et al. "Diffusion models for adversarial purification." ICML 2022.

[11] Sinha, Aman, et al. "Certifying some distributional robustness with principled adversarial training." ICLR 2018.

[12] Angelopoulos, Anastasios N., et al. "Conformal risk control." ICLR 2024.

[13] Liu, Zhuang, et al. "Rethinking the value of network pruning." ICLR 2019.

[14] Zhu, Michael, and Suyog Gupta. "To prune, or not to prune: exploring the efficacy of pruning for model compression." ICLR 2018.

[15] Jia, Ruoxi, et al. "Towards efficient data valuation based on the shapley value." PMLR 2019.

[16] Weber, Maurice G., et al. "Certifying out-of-domain generalization for blackbox functions." ICML 2022.

[17] Ruoss, Anian, et al. "Learning certified individually fair representations." NeurIPS 2020.

[18] Roh, Yuji, et al. "Sample selection for fair and robust training." NeurIPS 2021.