Teach AI What It Doesn't Know

Xuefeng Du

College of Computing and Data Science, Nanyang Technological University (NTU) Singapore,
xuefeng.du@ntu.edu.sg, https://d12306.github.io/

Talk Title. Teach AI What It Doesn't Know

Abstract. This talk surveys my research journey toward building reliable machine learning systems that behave safely and predictably in the open world. While modern machine learning models—including foundation models (FMs)—have demonstrated unprecedented capabilities, they often suffer from reliability failures under distribution shift, leading to overconfident mispredictions, hallucinated generations, or susceptibility to adversarial prompts. My research rethinks reliability not as an afterthought, but as a first-class algorithmic principle, to be optimized alongside accuracy with minimal human supervision.

The talk is organized around three key threads. To respect the allotted 20-30 minutes, the first and second parts will be **briefly** discussed.

1. Unknown-Aware Learning via Outlier Synthesis. I introduce a class of learning algorithms that synthesize "virtual outliers" in representation or pixel space to explicitly teach models what they don't know. This includes the VOS (Du et al. 2022), NPOS (Tao et al. 2023), and Dream-OOD (Du et al. 2023) frameworks, which shape the energy landscape around in-distribution data to avoid overconfidence on OOD.

2. Learning in the Wild with Unlabeled Data. I present theoretical insights and practical algorithms for leveraging unlabeled in-the-wild data to improve reliability. This includes SAL framework (Du et al. 2024a), which uses a gradient-based spectral method to separate potential outliers, and SCONE (Bai et al. 2023), which handles semantic and covariate shifts via constrained optimization. These results turn unlabeled data contamination into a learning signal.

3. Reliable Foundation Models. I explore reliability failures in LLMs and multimodal systems. I introduce HaloScope (Du, Xiao, and Li 2024) for hallucination detection via subspace separation on LLM representations, and TSV (Park et al. 2025) that performs LLM latent steering for improved hallucination detection. I will also briefly cover the LLM security and alignment, which includes VLMGuard (Du et al. 2024b) for detecting malicious prompts in vision-language models and a data-centric paradigm for AI alignment through source-aware feedback cleaning (Yeh et al. 2024).

Throughout the talk, I highlight how representation learning, data generation, and theoretical guarantees intersect to produce scalable, label-efficient reliability methods. I will also reflect on my broader vision: designing proactive and collaborative AI systems that anticipate uncertainty and support rich human-AI interaction—especially for underrepresented communities and emerging scientific domains.

This talk will be accessible to a broad AAAI audience, combining foundational algorithmic insights with real-world applications and forward-looking perspectives on the future of responsible AI.

References

Bai, H.; Canal, G.; Du, X.; Kwon, J.; Nowak, R. D.; and Li, Y. 2023. Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection. In *International Conference on Machine Learning*.

Du, X.; Fang, Z.; Diakonikolas, I.; and Li, Y. 2024a. How Does Unlabeled Data Provably Help Out-of-Distribution Detection? In *International Conference on Learning Representations*.

Du, X.; Ghosh, R.; Sim, R.; Salem, A.; Carvalho, V.; Lawton, E.; Li, Y.; and Stokes, J. W. 2024b. VLMGuard: Defending VLMs against Malicious Prompts via Unlabeled Data. *arXiv preprint arXiv:2410.00296*.

Du, X.; Sun, Y.; Zhu, X.; and Li, Y. 2023. Dream the Impossible: Outlier Imagination with Diffusion Models. In *Advances in Neural Information Processing Systems*.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations*.

Du, X.; Xiao, C.; and Li, Y. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *Advances in Neural Information Processing Systems*.

Park, S.; Du, X.; Yeh, M.-H.; Wang, H.; and Li, Y. 2025. Steer LLM Latents for Hallucination Detection. *International Conference on Machine Learning*.

Tao, L.; Du, X.; Zhu, X.; and Li, Y. 2023. Non-Parametric Outlier Synthesis. In *International Conference on Learning Representations*.

Yeh, M.-H.; Wang, J.; Du, X.; Park, S.; Tao, L.; Im, S.; and Li, Y. 2024. Challenges and Future Directions of Data-Centric AI Alignment. *International Conference on Machine Learning*.