# BOXUAN ZHANG

☎ +1-732-322-6932 | ✉ boxuan.zhang@rutgers.edu | Homepage

Linkedin | Github | Google Scholar

Rutgers University, New Brunswick, NJ - 08901

## Research Interests

My research centers on reliable machine learning as a cornerstone for trustworthy AI. I develop methods to improve model robustness (e.g., out-of-distribution detection), calibration (e.g., uncertainty quantification and hallucination), and behavioral consistency (e.g., frontier risk of self-replication), with a dual focus on addressing safety challenges of large language models and tackling critical issues in interdisciplinary domains (e.g., remote sensing and healthcare).

## Education

- **Rutgers University** *2025.09 - Present*
  *Ph.D. in CS, Department of Computer Science* Advisor: Prof. Ruixiang (Ryan) Tang
- **Wuhan University** *2022.09 - 2024.06*
  *M.Eng in AI, School of Computer Science* Advisor: Prof. Zengmao Wang
- **Wuhan University** *2018.09 - 2022.06*
  *B.Eng in CS, School of Computer Science* Advisor: Prof. Jing Xiao

## Publications

* Indicates the Equal Contribution

- **ACL 2025 Findings [Link]: Boxuan Zhang** and Ruqi Zhang, "CoT-UQ: Improving Response-wise Uncertainty Quantification in LLMs with Chain-of-Thought".
- **NeurIPS 2024 [Link]: Boxuan Zhang***, Jianing Zhu*, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han, "What If the Input is Expanded in OOD Detection?".
- **IEEE GRSL 2024 [Link]: Boxuan Zhang**, Zengmao Wang, and Bo Du, "Boosting Semisupervised Object Detection in Remote-Sensing Images With Active Teaching".

## Preprints

* Indicates the Equal Contribution

- **ArXiv 2025 [Link]:** Zicong He*, **Boxuan Zhang***, Weihao Liu*, Ruixiang Tang, and Lu Cheng, "What Shapes a Creative Machine Mind? Comprehensively Benchmarking Creativity in Foundation Models".
- **ArXiv 2025 [Link]: Boxuan Zhang***, Yi Yu*, Jiaxuan Guo, and Jing Shao, "Dive into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents".
- **Technical Report 2025 [Link]:** Shanghai AI Lab: Xiaoyang Chen, Yunhao Chen, ..., **Boxuan Zhang**, ... [30+ authors], "Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report".
- **ArXiv 2025 [Link]:** Zicong He*, **Boxuan Zhang***, and Lu Cheng, "Shakespearean Sparks: The Dance of Hallucination and Creativity in LLMs' Decoding Layers".

## Professional Experience

- **Research Assistant, Shanghai Artificial Intelligence Laboratory** *2025.03 - 2025.09*
  *Project Core Contributor and Leader, Center for Safe & Trustworthy AI, Advisor: Dr. Yi Yu* [*Project Link*]
  - Research on *Evaluation of Frontier Risk - Self-Replication Risk in LLM Agents.*
  - Establish authentic production environments and realistic tasks, including dynamic load balancing, and service maintenance under termination threats.
  - Propose fine-grained evaluation metrics, *Overuse Rate* (OR), *Aggregate Overuse Count* (AOC), and Risk Score ($\Phi_R$), to precisely quantify the frequency and severity of uncontrolled self-replication risks.
  - Participate in the SafeWork-F1 project as a core contributor, leading the section of self-replication risk. Submit one paper as co-first author to ICLR 2026, currently under review.
- **Research Intern, University of Illinois Chicago** *2024.11 - 2025.09*
  *Project Co-Leader, Responsible and Reliable AI Lab (R2 Lab), Advisor: Prof. Lu Cheng* [*Project Link 1* & *Link 2*]
  - Research on *Benchmarking Creativity in Foundation Models and Exploring its Interplay with Hallucination.*

- Propose a narrow definition of creativity tailored to LLMs and introduce HCL framework to quantify **H**allucination and **C**reativity across different **L**ayers of LLMs during decoding.
- Benchmark two complementary forms of Creativity - convergent creativity (tasks with constrained solutions like code generation) and divergent creativity (open-ended tasks like storytelling) with metrics "Usefulness, Originality, Surprise (U-O-S)" triplet derived from social science theories.
- Co-mentored a junior research intern, providing guidance on experiment design and paper writing. Submit two papers as co-first author to ARR 2025 and ICLR 2026, currently under review.

- **Research Intern, Purdue University** *2024.06 - 2025.02*
  *Project Leader, RZ-Lab, Advisor: Prof. Ruqi Zhang* [*Project Link*]
  - Research on *Uncertainty Quantification and Calibration in Large Language Models.*
  - Propose to quantify response-wise uncertainty by integrating LLMs' inherent reasoning capabilities through Chain-of-Thought (CoT) into the UQ process.
  - The proposed CoT-UQ achieves an average improvement of 5.9% AUROC compared to baselines.
  - Submit one paper to ACL 2025 (Accepted).
- **Research Intern, Hong Kong Baptist University** *2023.11 - 2024.06*
  *Project Leader, TMLR Group, Advisor: Prof. Bo Han and Dr. Jianing Zhu* [*Project Link*]
  - Research on *Out-of-Distribution (OOD) Detection for Reliable ML Model Deployment.*
  - Propose a novel perspective to employ different common corruptions on the input space to expand the representation dimension for OOD detection.
  - With the expectation among multiple input dimensions, our method performs a better ID-OOD separability.
  - Submit one paper as co-first author to NeurIPS 2024 (Accepted).
- **Research Intern, Wuhan University** *2023.08 - 2023.11*
  *Project Core Contributor, School of Civil Engineering, Advisor: Prof. Xiaoping Zhang* [*Project Link*]
  - Research on *Machine Learning for Tunnel Boring Machine (TBM) Excavation.*
  - Design an algorithm for accurate rock mass classification based on multi-feature optimization and efficient TBM parameter prediction using low-dimensional inputs.
  - This will help TBM operators to predict geological conditions in advance and the optimal operational parameters under geological variations.
  - Complete a technical paper and win the national third prize in the Second TBM Excavation Parameter Data Sharing and Machine Learning Competition.
- **Research Assistant, Wuhan University** *2022.11 - 2023.08*
  *Project Leader, SIGMA Group, Advisor: Prof. Zengmao Wang and Prof. Bo Du* [*Project Link*]
  - Research on *Active Learning for Semi-Supervised Object Detection in Remote Sensing Images*.
  - Propose to boost semi-supervised object detection with active teaching (SSOD-AT) in remote sensing images.
  - SSOD-AT can achieve high detection accuracy only with limited labeled samples, which helps to alleviate the dependency on limited labeled images in remote sensing scenarios.
  - Submit one paper to IEEE Geoscience and Remote Sensing Letters (Accepted).

## Teaching and Services

- **Teaching Assistant, Rutgers University** *2025.09 - Present*
  - CS 344: Design and Analysis of Computer Algorithms
- **Undergraduate Student Mentor, Wuhan University** *2022.09 - 2023.06*
  - Facilitated freshmen's transition to university life at School of Computer Science.
- **Journal Reviewer**
  - ISPRS Journal of Photogrammetry and Remote Sensing
- **Conference Reviewer**
  - NeurIPS (2025), ICLR (2026)

## Skills

**Programming:** Python (Pytorch), C/C++, Linux, Git, LaTeX

**Languages:** English (professional working proficiency), Chinese (native proficiency)