

# YIBIN WANG

✉ [wyb896409234@gmail.com](mailto:wyb896409234@gmail.com) ·  [Google Scholar](#) ·  [yibinwang.netlify.app](https://yibinwang.netlify.app) ·  [Github](#)

**Research Interests:** My research interests focus on **trustworthy AI**, particularly in the areas of adversarial robustness, uncertainty estimation and quantification.

## EDUCATION

**Huazhong University of Science and Technology (HUST)** *Sept. 2019 – June. 2024*  
B.E. in Computer Science (CS) (Excellent Class) , GPA: 3.82/4.00 - [Transcripts](#)

## EXPERIENCE

**Mitigating the Hallucination for LLM** *June. 2024 – Present*  
*Research Intern* | University of Illinois Urbana-Champaign (UIUC) *Advised by* [Prof. Huan Zhang](#)

**Generalization and Robustness of LLM** *Sept. 2023 – May. 2024*  
*Remote Research Intern* | Rutgers Machine Learning Lab, Rutgers University *Advised by* [Prof. Hao Wang](#)

- Conducted extensive research on adversarial robustness of large language models.
- Conducted in-depth research on Bayesian algorithms and their applications in large language models.

**Certified Adversarial Robustness in NLP** *Sept. 2021 – Aug. 2023*  
*Research Intern* | [John Hopcroft Lab for Data Science, HUST](#) *Advised by* [Prof. Kun He](#)

- Conducted extensive research on adversarial attack and defense in machine learning
- Conducted in-depth research on certified robustness based on convex relaxation

## PUBLICATIONS

\* indicates equal contribution

**BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models** *Preprint, under review*

- **Yibin Wang\***, Haizhou Shi\*, Ligong Han, Dimitris Metaxas, Hao Wang
- We introduce a principled Bayesian framework for improving large language models' generalization and uncertainty estimation. I contributed to the design of the algorithm and the writing of the paper, independently optimized the algorithm, implemented the code, and conducted the primary experiments.

**Continual Learning of Large Language Models: A Comprehensive Survey** *Preprint, under review*

- Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, **Yibin Wang**, Hao Wang
- Responsible for writing the parts related to large language models in the Preliminaries section.

**Robustness-Aware Word Embedding Improves Certified Robustness to Adversarial Word Substitutions** *In Findings of ACL 2023*

- **Yibin Wang\***, Yichen Yang\*, Di He, Kun He
- We transform the optimization problem of the model's certified robustness into an optimization problem of word embeddings through theoretical proofs. I independently complete all coding, experiments, and the main part of the paper writing.

## SURVICE

- Reviewer for NeurIPS 2024, EMNLP 2024

## ♡ AWARDS

---

*Honorable Mention*, Award on Mathematical Contest In Modeling

May. 2022

## ❖ MISCELLANEOUS

---

- Languages: English - IELTS overall score 7.0