Yibin Wang

yibinlwang@gmail.com · Google Scholar · yibinwang.netlify.app · Github

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

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

I got injured and took a one-year leave of absence from school in 2019.

Experience

Generalization, Calibration and Robustness of LLM *June. 2024 – June. 2025*

Research Intern | University of Illinois Urbana-Champaign (UIUC) *Advised by* Prof. Huan Zhang

Generalization, Calibration and Robustness of LLM *Sept. 2023 – May. 2024*

Remote Research Intern | Rutgers Machine Learning Lab, Rutgers University *Advised by* Prof. Hao Wang

Certified Adversarial Robustness in NLP *Sept. 2021 – Aug. 2023*

Research Intern | John Hopcroft Lab for Data Science, HUST *Advised by* Prof. Kun He

Publications

$*$ indicates equal contribution

Robustness-Aware Word Embedding Improves Certified Robustness to Adversarial Word Substitutions 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.

Continual Learning of Large Language Models: A Comprehensive Survey ACM Computing Surveys

- Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, **Yibin Wang**, Zifeng Wang, Sayna Ebrahimi, Hao Wang

BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models NeurIPS 2024

- **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 during fine-tuning. 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.

Training-Free Bayesianization for Low-Rank Adapters of Large Language Models NeurIPS 2025

- Haizhou Shi*, **Yibin Wang***, Ligong Han, Huan Zhang, Hao Wang
- We propose a training-free Bayesian framework to enhance uncertainty estimation and generalization of fine-tuned large language models in a computationally efficient way. I contributed to the design of the algorithm and implemented parts of the code and experiments.

Improving Data Efficiency for LLM Reinforcement Fine-tuning Through Difficulty-targeted Online Data Selection and Rollout Replay NeurIPS 2025

- Yifan Sun*, Jingyan Shen*, **Yibin Wang***, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, Huan Zhang

Efficient Uncertainty Estimation via Distillation of Bayesian Large Language Models In submission

- Harshil Vejendla*, Haizhou Shi*, **Yibin Wang**, Tunyu Zhang, Huan Zhang, Hao Wang

Token-Level Uncertainty Estimation for Large Language Model Reasoning In submission

- Tunyu Zhang*, Haizhou Shi*, **Yibin Wang**, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, Hao Wang

ℹ Service

- Reviewer for NeurIPS, ICLR, ACL, EMNLP