Ziran Yang

ziranyang0@gmail.com | website

Education

Peking University Sep 2021 - June 2025(Expected)

Bachelor of Engineering in Artificial Intelligence (Tong Class, Yuanpei College) *PKU*

- **GPA(cumulative):** 3.73/4.00, Ranks 4/32 in Tong Class (A Special Pivot AI Program led by Song-Chun Zhu).
- **Core Courses:** Mathematical Foundation for AI (93.5), Computer Vision (90.6), Intro to Computer Systems (92), Discrete Mathematics and Structures (89), Directed Research in AI (92).
- **Standard Tests:** TOEFL iBT: 105 (Speaking 24).
- **Selected Honors and Awards:**
  * 2024: Yuanpei College Undergraduate Research Award and Academic Star Title
  * 2024: SenseTime Scholarship Nomination Award
  * 2024: Song Qingling Future Scholarship
  * 2024: Fifth Yuanpei Young Scholar Award
  * 2023: Peking University Institute for Artificial Intelligence Annual Technology Day: Best Innovation Award
  * 2023: Peking University Learning Excellence Award
  * 2023: Shu Qi Scholarship, Peking University
  * 2022: Peking University Learning Excellence Award
  * 2022: Lee Wai Wing Scholarship, Peking University
  * 2021: Peking University Freshman Scholarship
  * 2021: National College Entrance Examination (aka "Gaokao", Anhui Provincial), Ranking 11/230000+
  * 2020: Chinese Mathematics Olympiad (Anhui Provincial), First Prize
  * 2019: Ministry of Education Talent Program Annual Outstanding Thesis

Experience

MixLab, UC San Diego Halicioglu Data Science Institute Apr 2024 – Present

Visiting Research Intern *Advisor: Prof. Zhiting Hu*

- Understanding the sources of uncertainty in large models and how to quantify them through a unified decomposition framework to enhance reliability and enable effective error detection for safer deployment.

LangAI Nexus Lab, Stony Brook CS Department Apr 2024 – Present

Remote Research Intern *Advisor: Prof. Jiawei Zhou*

- Enhancing large vision-language models by quantifying visual input uncertainty and applying uncertainty-guided token dropout to reduce hallucinations and improve output reliability.

PAIR Lab, Peking University May 2023 – Present

Research Intern *Advisor: Prof. Yaodong Yang*

- Exploring how multi-agent dynamics and game theory can be leveraged in large language model settings to develop dynamic red-team strategies that mitigate mode collapse and enhance the safety and reliability of LLMs.

Tong Class, Peking University Sep 2021 – Present

Undergraduate Student *Advisor: Prof. Yixin Zhu, Prof. Song-Chun Zhu*

Publications and Manuscripts

* indicates equal contributions

Evolving Diverse Red-team Language Models in Multi-round Multi-agent Games

Chengdong Ma, **Ziran Yang***, Hai Ci, Jun Gao, Minquan Gao, Xuehai Pan, Yaodong Yang* *Under Review in T-PAMI*

Understanding the Sources of Uncertainty for Large Language and Multimodal Models

***Ziran Yang**, Shibo Hao, Hao Sun, Lai Jiang, Qiyue Gao, Binglin Zhou, Yian Ma, Zhiting Hu* *Preprint*

From Uncertainty to Trust: Enhancing Reliability in Vision-Language Models with Uncertainty-Guided Context Dropout

Yixiong Fang,**Ziran Yang***, Zhaorun Chen, Zhuokai Zhao, Jiawei Zhou* *Under Review in CVPR*

Panacea: Pareto Alignment via Preference Adaptation for LLMs

Yifan Zhong, Chengdong Ma*, Xiaoyuan Zhang*, **Ziran Yang**, Qingfu Zhang, Siyuan Qi, Yaodong Yang* *NeurIPS 2024*

SafeSora: Towards Safety Alignment of Text2Video Generation via a Human Preference Dataset

*Josef Dai, Tianle Chen, Xuyao Wang, **Ziran Yang**, Taiye Chen, Jiaming Ji, Yaodong Yang* *NeurIPS DB Track 2024*