YIGE YUAN

State Key Laboratory of AI Safety
Institute of Computing Technology, Chinese Academy of Sciences
No.6 Kexueyuan South Road Zhongguancun, Haidian District, Beijing, China

yuanyige20z@ict.ac.cn
github.com/yuanyige
Google Scholar

RESEARCH INTEREST

My research goal is to build **Trustworthy AI** that performs reliably across diverse scenarios. To achieve this, I worked on generalization, alignment and reasoning/planning across the domains of graph, vision, and language. My research includes:

- **Machine Learning Generalization**: To make models generalize stably under domain shifts, noises, or perturbations.
- **Large Language Model Alignment**: To align large language models with human values and safety requirements.
- **AI System Reasoning and Planning**: To enhance AI's logical reasoning and strategic planning in complex environments.

EDUCATION

Institute of Computing Technology, Chinese Academy of Sciences Sep 2020 - current
Ph.D. in Computer Software and Theory (Advisor: Prof. Xueqi Cheng & A.P. Bingbing Xu)

Xidian University, School of Cyberspace Security Sep 2016 - Jun 2020
B.S. in Information Security (Experimental Class, GPA: 3.8/4.0)

INTERNSHIP

Tongyi Lab, Alibaba Group Feb 2025 - current
Research Internship in Large Language Models and Multi-Agent Systems

PUBLICATIONS

C: conference, J: journal, W: workshop, P: preprint / * equal contribution

Machine Learning Generalization

[C1] TEA: Test-time Energy Adaptation
Yige Yuan, Bingbing Xu, Liang Hou, Fei Sun, Huawei Shen, Xueqi Cheng
IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024, Main Track, CCF-A

[C2] PDE+: Enhancing Generalization via PDE with Adaptive Distributional Diffusion
Yige Yuan, Bingbing Xu, Bo Lin, Liang Hou, Fei Sun, Huawei Shen, Xueqi Cheng
AAAI Conference on Artificial Intelligence (**AAAI**), 2024, Main Track, CCF-A

[J1] Towards Generalizable Graph Contrastive Learning: An Information Theory Perspective
Yige Yuan, Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, Wen Zheng, Xueqi Cheng
Neural Networks (**NN**), Volume 172, CCF-B, Q1, IF=8.4

[P1] MITA: Bridging the Gap between Model and Data for Test-Time Adaptation
Yige Yuan, Xu Bingbing, Liang Hou, Fei Sun, Huawei Shen, Xueqi Cheng
IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), UnderReview

[C3] Augmentation-Aware Self-Supervision for Data-Efficient GAN Training
Liang Hou, Qi Cao, **Yige Yuan**, Songtao Zhao, Chongyang Ma, Siyuan Pan, et al.
Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2023, Main Track, CCF-A

[C5] InfoNCE is a Free Lunch for Semantically guided Graph Contrastive Learning
Zixu Wang, Bingbing Xu, **Yige Yuan**, Huawei Shen and Xueqi Cheng
International ACM Conference on Research and Development in Information Retrieval (**SIGIR**), 2025, Full Paper, CCF-A

[C4] Negative as Positive: Enhancing Out-of-distribution Generalization for Graph Contrastive Learning
Zixu Wang, Bingbing Xu, **Yige Yuan**, Huawei Shen and Xueqi Cheng
International ACM Conference on Research and Development in Information Retrieval (**SIGIR**), 2024, Short Paper, CCF-A

[C6] History Driven Sampling for Scalable Graph Neural Networks
Yang Li, Bingbing Xu, Fei Sun, Qi Cao, **Yige Yuan**, and Huawei Shen
International Conference on Database Systems for Advanced Applications (**DASFAA**), 2024, Reseach Track, CCF-B

[P2] MIGE: A Unified Framework for Multimodal Instruction-Based Image Generation and Editing
Xueyun Tian, Wei Li, Bingbing Xu, **Yige Yuan**, Yuanzhuo Wang, Huawei Shen
Annual Meeting of the Association for Computational Linguistics (**ACL**), 2025, UnderReview

Large Language Model Alignment

[C7] Inference-time Alignment in Continuous Space
Yige Yuan*, Teng Xiao*, Yunfan Li, Bingbing Xu, Shuchang Tao, Yunqi Qiu, Huawei Shen, Xueqi Cheng
International Conference on Learning Representations (**ICLR**), 2025, Bi-Align Workshop
Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2025, UnderReview

[C8] Fact-Level Calibration and Correction for Long-Form Generations
Yige Yuan, Xu Bingbing, Hexiang Tan, Fei Sun, Teng Xiao, Wei Li, Huawei Shen, Xueqi Cheng
International ACM Conference on Research and Development in Information Retrieval (**SIGIR**), 2025, Short Paper, CCF-A

[C9] SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters
Teng Xiao*, **Yige Yuan***, Zhengyu Chen, Mingxiao Li, Shangsong Liang, Zhaochun Ren, Vasant G Honavar
International Conference on Learning Representations (**ICLR**), 2025, Main Conference

[C10] On a Connection Between Imitation Learning and RLHF
Teng Xiao, **Yige Yuan**, Mingxiao Li, Zhengyu Chen, Vasant G Honavar
International Conference on Learning Representations (**ICLR**), 2025, Main Conference

[C11] Calibrated Preference Optimization for Direct Language Model Alignment
Teng Xiao, **Yige Yuan**, Huaisheng Zhu, Mingxiao Li, Vasant G Honavar
Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2024, Main Track, CCF-A

[C12] How to Leverage Demonstration Data in Alignment for Large Language Model? A Self-Imitation Learning Perspective
Teng Xiao, Mingxiao Li, **Yige Yuan**, Huaisheng Zhu, Chao Cui, Vasant G Honavar
Conference on Empirical Methods in Natural Language Processing (**EMNLP**), 2024, Main, CCF-B

[C13] Unveiling the Potential of LLMs in Simulated Society: A Knowledge-Driven LLM Agent Framework for User Modeling
Shengmao Zhu, Bingbing Xu, **Yige Yuan**, Bin Xie, Yunfan Li, Huawei Shen
ACM Web Conference (**WWW**), 2025, Companion Proceedings, CCF-A

[C14] Score Consistency Meets Preference Alignment: Dual-Consistency for Partial Reward Modeling
Bin Xie, Bingbing Xu, **Yige Yuan**, Shengmao Zhu, Huawei Shen
Annual Meeting of the Association for Computational Linguistics (**ACL**), 2025, Main, CCF-A

[P3] Learn over Past, Evolve for Future: Generating Future Behavior for Social Bot Detection
Xiao Zhang, **Yige Yuan**, Bingbing Xu, Huawei Shen
International Joint Conference on Artificial Intelligence (**IJCAI**), 2025, UnderReview

AI System Reasoning and Planning

[P4] Incentivizing Strong Reasoning from Weak Supervision
Yige Yuan*, Teng Xiao*, Shuchang Tao, Xue Wang, Jinyang Gao, Bolin Ding, Bingbing Xu
Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2025, UnderReview

HONORS & AWARDS

First Place, *AgentSociety Challenge @ WWW 2025*	2025
National Scholarship, *Ministry of Education of the People's Republic of China*	2024
First-Class Scholarship, *University of Chinese Academy of Sciences*	2024
Presidential Scholarship, *Institute of Computing Technology, Chinese Academy of Sciences*	2023
First-Class Scholarship, *University of Chinese Academy of Sciences*	2022
Outstanding Student Award, *University of Chinese Academy of Sciences*	2022
First Prize, *The 12th National College Students Information Security Contest*	2019
First Prize, *15th National Science and Technology Academic Competition of Challenge Cup*	2017

INVITED TALKS

WiseModel Talk, On a Connection Between Imitation Learning and RLHF	April 2025
NICE Webinar, On a Connection Between Imitation Learning and RLHF	March 2025
AITime Youth PhD Talk, On a Connection Between Imitation Learning and RLHF	March 2025
LOGS Webinar, Partial Differential Equation-Driven Generalizable Neural Networks	Mar 2024
AITime Webinar, TEA: Test-time Energy Adaptation	April 2024
WizSci Webinar, PDE+: Enhancing Generalization via PDE with Adaptive Distributional Diffusion	Jan 2024

ACADEMIC SERVICES

Conference Reviewer: NeurIPS (2024, 2025), ICML 2025, ICLR 2025, AISTATS 2025, KDD 2025, WWW 2025, ACMMM 2025, AAAI 2025, IJCAI 2025, ACL 2025, EMNLP 2024, COLING 2025, ACL Rolling Review, MIDL 2025, IJCNN 2025
Journal Reviewer: IEEE Transactions on Knowledge and Data Engineering (TKDE), Applied Intelligence (APIN), CAAI Transactions on Intelligence Technology, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)