YANG, YAO

YaoYangAcademia@outlook.com | (86) 150-3529-5658
https://evigbyen.github.io/ | Hong Kong SAR & Shanghai, P.R.China

Version 5.2 # Updated on Oct 13 2025

1. SUMMARY

Completed **MSc in Artificial Intelligence at HKU** and **BEng in Data Science and Big Data Technology at NEU**, YANG, YAO combines a solid theoretical grounding in mathematics and statistics with deep practical experience in algorithm programming. Thus far, he has led **three first-author papers** (two accepted by ACL 2025 and ACM Multimedia 2025, one under review), and has been credited as **a core contributor on a technical report**. As an early-stage researcher, his research interests focus on **(1) (Multimodal) Large Language Models**, **(2) Agents**, and **(3) Safe and Trustworthy AI**, while staying **OPEN to exploration across broader topics**. Currently, He is **pursuing PhD opportunities** for **Fall 2026** to further contribute to these domains.

2. EDUCATION

MSc in Artificial Intelligence, University of Hong Kong **Sep 2023 - Jun 2025**
GPA of 3.24/4.00, Faculty of Science Hong Kong SAR
► Core courses: Optimization Theory, Machine Learning, Computational Intelligence, and Deep Learning

BEng in Data Science and Big Data Technology, Northeastern University **Sep 2019 - Jun 2023**
GPA of 3.76/5.00 (87.6/100.0), School of Mathematics and Statistics Shenyang & Qinhuangdao
► Core courses: Advanced Algebra, Mathematical Analysis, Probability Theory, Bayes Statistics, Random Process, Python Programming, Distributed Clusters, MySQL Database, Operating System, and Computer Networks

3. PUBLICATIONS & SUBMISSIONS

6 additional papers not listed

[1] **Yang Yao**, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, Yingchun Wang. 2025. **A Mousetrap: Fooling Large Reasoning Models for Jailbreak with Chain of Iterative Chaos**. (**ACL 2025, Poster, Findings with OA 3.33/5 and Meta 3.5/5**)

[2] **Yang Yao**, Lingyu Li, Jiaxin Song, Chiyu Chen, Zhenqi He, Yixu Wang, Xin Wang, Tianle Gu, Jie Li, Yan Teng, Yingchun Wang. 2025. **Argus Inspection: Do Multimodal Large Language Models Possess the Eye of Panoptes?** (**ACM Multimedia 2025, Oral**)

[3] **Yang Yao (Core Contributor)**, et al. 2025. **SafeWork-R1: Coevolving Safety and Intelligence under the AI-45° Law**. (**Intern Safework Technical Report**)

[4] **Yang Yao**, Yixu Wang, Yuxuan Zhang, Yi Lu, Tianle Gu, Lingyu Li, Dingyi Zhao, Keming Wu, Haozhe Wang, Ping Nie, Yan Teng, Yingchun Wang. 2025. **A Rigorous Benchmark with Multidimensional Evaluation for Deep Research Agents: From Answers to Reports**. (Under Review)

[5] Lingyu Li, **Yang Yao**, Yixu Wang, Chunbo Li, Yan Teng, Yingchun Wang. 2025. **The Other Mind: How Language Models Exhibit Human Temporal Cognition**. (Under Review)

[6] Yixu Wang, Xin Wang, **Yang Yao**, Xinyuan Li, Yan Teng, Xingjun Ma, Yingchun Wang. 2025. **SafeEvalAgent: Toward Agentic and Self-Evolving Safety Evaluation of LLMs**. (Under Review)

4. ACADEMIC EXPERIENCES

Multimodal Value Alignment: Benchmark and Reward Model Fine-tuning **Nov 2024 - Mar 2025**
► Constructed benchmarks with Multimodal Large Language Models (MLLMs, e.g., GPT, Gemini, and Qwen) based on the scenarios and elements of real-world human ethical values to evaluate the responses of MLLMs in terms of value alignment on response quality, security, reliability, and compliance.

► Built and optimized two pipelines based on image generation with DALL·E 3 and search with Google Image Search, and generated about 200,000 multimodal value data layer by layer through prompts engineering.
► Designed Outcome Reward Model (ORM) by full and Low-Rank Adaptation (LoRA) Supervised Fine-Tuning (SFT) Qwen2.5-VLs (7B, and 72B) and Process Reward Model (PRM), with accuracy rates of 93.9% and 96.8%.

Jailbreak Large Reasoning Models with Chain of Iterative Injections **Jan - Feb 2025**
► Proposed the first jailbreak framework to Large Reasoning Models (LRMs) via flaws of reasoning process.
► Designed a novel component called chaos machine to transform jailbreak prompts with diverse one-to-one mappings (injections) which are iteratively generated and embedded into the reasoning chain to strengthen the variability and complexity and also promote more robust attacks.
► Trapped LRMs with mismatched generalization enhanced and more competing objectives with success rates of 96%, 86% and 98% respectively attacking o1-mini, Claude-sonnet, and Gemini-thinking on presented toxic dataset Trotter and 87.5%, 86.58% and 93.13% respectively attacking Claude-sonnet, well-known for its safety, on more mainstream benchmarks, AdvBench, StrongREJECT, and HarmBench.

Stochastic Optimization Method for AUC Maximization **Aug - Oct 2024**
► Explored the optimization methods for the Area Under receiver operating characteristic Curve (AUC) maximization based on the equivalent objectives, focusing on the fact that the classical stochastic optimization methods are not directly applicable to AUC maximization due to the nonlinearity of the objective functions.
► Designed diverse forms of alternative loss functions based on mathematical tricks (e.g., property of moment generating function, and Taylor expansion), and verified the convergence and stability of the algorithms.
► Achieved AUC scores of 0.9296 and 0.7197 respectively on the simple and difficult benchmarks of imbalanced classification tasks, surpassing the baseline.

5. INTERNSHIPS

Shanghai Artificial Intelligence Laboratory, Safety and Trustworthiness AI Center **Nov 2024 - Present**
Large Language Models Research Intern (Safety and Value) Shanghai
► Constructed multimodal value benchmarks using MLLM APIs based on image generation and image search, and trained 7B and 72B outcome reward models by full and LoRA SFT.

China Eastern Airlines Co., Ltd., Frontier Technology Application Research Dept. **Sep - Oct 2024**
Artificial Intelligence Algorithm Engineer (Deep Learning) Shanghai
► Pre-processed and annotated image data in VOC format and developed a fuel oil leakage detection model for aircraft maintenance using YOLOX, demonstrating its feasibility and achieving robust performance in applications.

6. PROFESSIONAL SKILLS

Python (PyTorch, TensorFlow, NumPy, Pandas, LLMs API, Matplotlib, Scikits-Learn), **R**, Julia, **MySQL**, Apache Hadoop Distributed Clusters (MapReduce & Zookeeper & Hive & HBase & Spark & Flume & Flink & Kafka), etc.

7. LEADERSHIP & TEAMWORK EXPERENCES

Student Assistant Principal, Northeastern University at Qinhuangdao Dec 2020 - Dec 2021
Coach of Debate Team, Sydney Smart Technology College, Northeastern University Sep 2020 - Feb 2021

8. QUALIFICATIONS & PRIZES & COMPETITIONS

Senior Big Data Application Engineer Certificate, Talent Training Project, Education and Examination Center, Ministry of Industry and Information Technology, P.R.China Mar 2022
University Scholarships (4 times) & Innovation and Entrepreneurship Scholarships (2 times) Undergraduate
Honourable Mention, Mathematical and Interdisciplinary Contest in Modeling (MCM-ICM) May 2022
Third Prize, Chinese Mathematics Competitions (CMC) Dec 2020