Yougang Lyu

+31 0623470832 | youganglyu@gmail.com | Homepage | Google Scholar

Summary

- **PhD candidate at the University of Amsterdam** (US News: 33rd, QS: 55th), graduating in July 2025.
- **Extensive research internship experience with a proven ability to translate complex industrial challenges into AI solutions**.
- **Specializes in pioneering AI research**—including **general LLM alignment**, **development of domain-specific LLMs** and **creation of autonomous, personalized, and safe web agents**.
- **Authored 11 top-tier IR and NLP papers** (ICLR, WWW, AAAI, EMNLP, Recsys, IPM), **with 8 as first author**, and recipient of the **Recsys24 Best Paper Award**.

Education

- **University of Amsterdam** *Jun 2021 - July 2025*
  PhD, Information Institution Amsterdam, Netherland
  - Research Team: Information Retrieval Lab
  - Supervisor: Prof. Maarten de Rijke and Assoc. Prof. Zhaochun Ren
  - Thesis: Aligning Language Models with Human Understanding and Behavior
- **Shandong University** *Sep 2017 - Jun 2021*
  Bachelor, School of Computer Science and Technology Qingdao, China
  - Research Team: Information Retrieval Lab
  - Supervisor: Assoc. Prof. Zhaochun Ren
  - Thesis: Legal Judgment Prediction Based on Criminal Elements

Working Experience

- **MBZUAI** *Feb 2025 - Present*
  Research Internship, NLP Department Abu Dhabi, UAE
  - Main duties: I am exploring the limitations of existing deep research agents (OpenAI, Gemini, and Perplexity) in accomplishing complex user web shopping queries and developing better web shopping agents.
- **Baidu Inc.** *Jun 2023 - Feb 2025*
  Research Internship, Search Science Team Beijing, China
  - Main duties: I proposed knowledge-aware fine-tuning and multi-agent contrastive alignment methods to solve strong-to-weak alignment and weak-to-strong alignment, respectively.
  - Achievements: EMNLP 24 and ICLR 25.
- **Baidu Inc.** *Jun 2021 - Feb 2022*
  Research Internship, Search Science Team Beijing, China
  - Main duties: I devised a debiased natural language understanding method, which reduces biased features and improves the performance of natural language understanding. The work was accepted by AAAI23.
  - Achievements: AAAI23.

Selected Publications

- **Feature-Level Debiased Natural Language Understanding (AAAI 23)**
  Yougang Lyu, *Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, Zhaochun Ren*
  - Proposed debiasing contrastive learning to address limitations of existing NLU debiasing methods by integrating debiased positive sampling and dynamic negative sampling, reducing biased latent representations and adapting to dynamic bias patterns.
  - Achieved state-of-the-art out-of-distribution performance on three NLU benchmarks while maintaining in-distribution accuracy.
- **KnowTuning: Knowledge-aware Fine-tuning for Large Language Models (EMNLP '24)**
  Yougang Lyu, *Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, Zhaochun Ren*

- Proposed *KnowTuning*, a fine-tuning framework that enhances factuality in LLMs by improving fine- and coarse-grained knowledge awareness.
- Achieved state-of-the-art performance on general and medical QA tasks across various LLMs, validated by automatic and human evaluations.

- **MACPO: Weak-to-Strong Alignment via Multi-Agent Contrastive Preference Optimization (ICLR '25)**
  Yougang Lyu, *Lingyong Yan, Zihan Wang, Dawei Yin, Pengjie Ren, Maarten de Rijke, Zhaochun Ren*
  - Proposed a multi-agent framework for aligning strong LLMs via weak supervision, tackling the weak-to-strong alignment challenge.
  - Achieved state-of-the-art results on HH-RLHF and PKU-SafeRLHF, improving both student and teacher models through iterative optimization.

- **Cognitive Debiasing Large Language Models for Decision-Making (Under Review)**
  Yougang Lyu, *Shijie Ren, Yue Feng, Zihan Wang, Zhumin Chen, Zhaochun Ren, Maarten de Rijke*
  - Developed a self-debiasing framework that iteratively refines prompts to reduce cognitive bias in LLM-based decision-making.
  - Outperformed existing debiasing and prompt engineering methods on finance, healthcare, and legal benchmarks.

- **Improving Legal Judgment Prediction through Reinforced Criminal Element Extraction (IPM '23)**
  Yougang Lyu, *Zihan Wang, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Xiaozhong Liu, Yujun Li, Hongsong Li, Hongye Song*
  - Proposed Criminal Element Extraction Network, a reinforcement learning-based framework for improving Legal Judgment Prediction by extracting four types of key criminal elements: criminal, target, intentionality, and criminal behavior.
  - Achieved significant performance gains on real-world LJP datasets, demonstrating that explicit criminal element modeling enhances the prediction of law articles, charges, and penalties.

- **Multi-Defendant Legal Judgment Prediction via Hierarchical Reasoning (EMNLP 23)**
  Yougang Lyu, *Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, Zhaochun Ren*
  - Proposed the Hierarchical Reasoning Network, which models multi-defendant judicial processes as hierarchical reasoning chains to predict criminal relationships, sentencing circumstances, law articles, charges, and penalties for each defendant.
  - Constructed and released *MultiLJP*, the first real-world dataset for multi-defendant LJP. Experiments on MultiLJP demonstrate the effectiveness of HRN in generating accurate, defendant-specific judgment results.

Awards

- ACM RecSys Best Full Paper Award, 2024
- Student Travel Award in AAAI, 2023
- China National Scholarship, 2023
- Dean scholarship, Shandong University, 2021
- First-Class Prize in Chinese Chemistry Olympiad, 2016

Presentation and Academic Services

- **Conference talks:** AAAI 2023, EMNLP 2023/2024, Recsys 2024
- **Invited talk:** University of Glasgow, 2025
- **Workshop Organizer:** The 1st Workshop on Human-Centered Recommender Systems at WWW, 2025
- **Reviewer:** PC Member of NeurIPS, ACL, EMNLP, SIGIR, CIKM, ECIR, TOIS and TKDE

Others

- **Programming Languages:** Python, Latex, PyTorch
- **Teaching Experience:** Teaching assistant at University of Amsterdam for RecSys course
- **Language:** English (Fluent), Mandarin (Mother tongue)