Zhenrui Yue

Homepage | LinkedIn | GitHub | Google Scholar
zhenrui3@illinois.edu | +1 650-605-5109 | Champaign, IL 61820

Education

University of Illinois Urbana-Champaign Aug 2021 – Present
Ph.D. in Cognitive Science & Language Processing
- **Advisor**: Dr. Dong Wang; **Research**: Language Models, Recommender Systems & Information Retrieval
- **Thesis**: Advancing Language Models for Knowledge- and Reasoning-Intensive Tasks: From Training to Inference

Technische Universität München Apr 2019 – May 2021
M.Sc. in Robotics, Cognition, Intelligence
- **Thesis** (ETH Zurich w/ Dr. Stefan Feuerriegel): Question Answering under Domain Shift
- **Exchange** (UC San Diego w/ Dr. Julian McAuley): Supported by Max Weber Scholarship from 2019 to 2020

Technische Universität München Oct 2015 – Mar 2019
B.Sc. in Mechanical Engineering
- **Track**: Mechatronics and Information Technology
- **Courses**: Mathematics, Mechanics, Control Theory, Simulation, Intro to CS, Software Eng., Algo & Data Struct. etc.

Experience

Meta GenAI May 2025 – Present
Research Intern
- Work with the safety post-training team on different reinforcement learning paradigms across diverse tasks
- Design continuous space post-training methods for LLMs with enhanced reasoning performance and efficiency

Google DeepMind May 2024 – Dec 2024
Student Researcher
- Explored the boundaries of long-context LLMs, designed inference strategies accordingly and evaluated the outcome
- Introduced a novel test-time scaling paradigm for long-context RAG that achieves substantial gains on knowledge intensive tasks; the proposed methods and findings are published in a first-author paper at ICLR 2025

NVIDIA May 2023 – Aug 2023
Applied Research Intern
- Investigated LLM-based recommendation methods and evaluated the performance of existing approaches
- Designed and implemented a two-stage LLM-based recommendation framework for efficient item retrieval and personalized ranking; the analysis and results are presented in the PGAI workshop co-located with CIKM 2023

Selected Publications

Hybrid Latent Reasoning via Reinforcement Learning
Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, Dong Wang
In Submission, 2025

Retrieval Augmented Conversational Recommendation with Preference Optimization
Zhenrui Yue, Honglei Zhuang, Zhen Qin, Zhankui He, Huimin Zeng, Julian McAuley, Dong Wang
In Submission, 2025

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning
Bowen Jin, Hansi Zeng, **Zhenrui Yue**, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, Jiawei Han
COLM 2025 [Paper] [Code]

Preference-Optimized Retrieval and Ranking for Efficient Multimodal Recommendation
Zhenrui Yue, Huimin Zeng, Yueqi Wang, Julian McAuley, Dong Wang
KDD 2025 [Paper] [Code]

Inference Scaling for Long-Context Retrieval Augmented Generation
Zhenrui Yue*, Honglei Zhuang*, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, Michael Bendersky
ICLR 2025 [Paper]

Train Once, Deploy Anywhere: Matryoshka Representation Learning for Multimodal Recommendation
Yueqi Wang*, **Zhenrui Yue***, Huimin Zeng, Dong Wang, Julian McAuley
Findings of EMNLP 2024 [Paper] [Code]

Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments
Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, Dong Wang
ACL 2024 [Paper] [Code]

Linear Recurrent Units for Sequential Recommendation
Zhenrui Yue*, Yueqi Wang*, Zhankui He, Huimin Zeng, Julian McAuley, Dong Wang
WSDM 2024 [Paper] [Code]

LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking
Zhenrui Yue, Sara Rabhi, Gabriel Moreira, Dong Wang, Even Oldridge
PGAI@CIKM 2023 [Paper] [Code]

Zero- and Few-Shot Event Detection via Prompt-Based Meta Learning
Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, Dong Wang
ACL 2023 [Paper] [Code]

On Attacking Out-Domain Uncertainty Estimation in Deep Neural Networks
Huimin Zeng, **Zhenrui Yue**, Yang Zhang, Ziyi Kou, Lanyu Shang, Dong Wang
IJCAI 2022 [Paper] [Code]

Defending Substitution-Based Profile Pollution Attacks on Sequential Recommenders
Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, Dong Wang
RecSys 2022 [Paper] [Code]

Contrastive Domain Adaptation for Question Answering using Limited Text Corpora
Zhenrui Yue, Bernhard Kratzwald, Stefan Feuerriegel
EMNLP 2021 [Paper] [Code]

Black-Box Attacks on Sequential Recommenders via Data-Free Model Extraction
Zhenrui Yue*, Zhankui He*, Huimin Zeng, Julian McAuley
RecSys 2021 [Paper] [Code]

Awards

Shiyu Scholarship	2024
SIGIR & RecSys Travel Award	2023
Bosch AIoT Scholarship	2021
E-fellows.net Scholarship	2020
Max Weber Academic Exchange Scholarship	2019
Audi China Scholarship	2018
Max Weber Program	2017

Teaching

Computer Networks	Spring 2024
Concepts of Machine Learning	Fall 2023
Introduction to Database	Fall 2022
Analytical Fndts Info Problems	Spring 2022

Talks

Cohere Labs: Inference Scaling for Long-Context Retrieval Augmented Generation	Jan 2025
LinkedIn: LLM for Efficient Sequential Recommendation	Aug 2024
Cohere: Linear Recurrent Units for Recommendation	Jul 2024

Miscellaneous

Coding: Python, C / C++; experiences with PyTorch, Jax, TensorFlow, Docker, Git, Bash, AWS, etc.
Languages: Native in Mandarin and Cantonese, proficient in English and German, beginner in Spanish
Service: AAAI, ARR (ACL, EMNLP, NAACL, etc.), COLING, COLM, ICCV, ICLR, KDD, NeurIPS, SIGIR, WWW, etc.