Wanli Yang

Curriculum Vitae

(+86) 18759861072
yangyywl@gmail.com
WanliYoung
yangwl.site
Google Scholar

Education

2024–Present **Institute of Computing Technology, Chinese Academy of Sciences**
Ph.D. Candidate in Cyberspace Security
- Research Advisor: Prof. Fei Sun & Prof. Xinran Liu

2020–2024 **Nankai University, College of Software**
B.E. in Software Engineering
- Rank: 5/140, CET4: 616, CET6: 614

Research Experience

12/2023 –Present **Reliable & Practical Evaluation of Model Editing**
- **Advisor**: Prof. Fei Sun
- **Institution**: Institute of Computing Technology, CAS
- Uncovering the potential of model editing to trigger LLMs collapse and proposing employing perplexity as a surrogate metric to monitor model collapse.
  Accepted to **Findings of ACL 2024**.
- Revealing existing model editing evaluation adopts inappropriate strategies, such as teacher forcing during testing, which substantially overestimate the effectiveness of existing techniques.
  Accepted to **Main Conference of ACL 2025**.

12/2022 –03/2023 **Graph Contrastive Learning for Recommendation**
- **Advisor**: Prof. Wayne Xin Zhao
- **Institution**: Gaoling School of Artificial Intelligence, RUC
- Contributing implementations of the latest recommendation models for **RecBole** (GitHub 3.3k Stars).

Internship Experience

11/2023 –04/2025 **Reasons Behind LLMs Collapse Caused by Editing**
- **Mentor**: Dr. Xinyu Ma
- **Company**: Baidu Inc.
- Revealing the root causes behind the model collapse triggered by editing and proposes a straightforward solution to prevent collapse and achieve remarkable editing performance.
- Accepted to **Findings of EMNLP 2024**.

04/2025 –Present **Effective Model Editing through Fine-tuning**
- **Mentor**: Dr. Hongyu Zang
- **Company**: Meituan Inc.
- Investigating the underlying reasons for the failure of fine-tuning in model editing and devising effective tuning strategies.
- Working on progress.

Publications and Preprints

ACL 2024 Findings — **Wanli Yang**, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, Xueqi Cheng. The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse.

ACL 2024 Main — Hexiang Tan, Fei Sun, **Wanli Yang**, Yuanzhuo Wang, Qi Cao, Xueqi Cheng. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?

EMNLP 2024 Findings — **Wanli Yang**, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, Huawei Shen. The Fall of ROME: Understanding the Collapse of LLMs in Model Editing.

ACL 2025 Main — **Wanli Yang**, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, Xueqi Cheng. The Mirage of Model Editing: Revisiting Evaluation in the Wild.

Skills

Programming — Python, C/C++, Java, SQL

Frameworks — PyTorch, TensorFlow, JAX

Others — Git, LaTeX, Linux

Honors and Awards

2024 — **Outstanding Graduates (Top 5%)** — Nankai University

2023 — **College Scholarship** — University of Chinese Academy of Sciences

2022 & 2023 — **Academic Excellence Scholarship (Top 5%)** — Nankai University

2021 & 2022 — **Arts and Sports Scholarship (Top 2%)** — Nankai University

2021 — **Innovation Scholarship (Top 3%)** — Nankai University

Hobbies

Sports — Fitness, Swimming, Cycling, Dragon Boat Racing (former paddler for the Dragon Boat Team of Nankai University), etc.

Reading — Works of Haruki Murakami.