

# Zeping Yu

Homepage: <https://zepingyu0512.github.io>

Google Scholar: <https://scholar.google.com/citations?user=OdpmpDsAAAAJ>

Target Position: Research Scientist / Applied Scientist (Large Language Models)

Email: zepingyu@foxmail.com

Mobile: +44-7529930611

Wechat: +86-18621601510

## EDUCATION

- |  |                                  |
|--|----------------------------------|
| • <b>University of Manchester</b>      | Manchester, UK                   |
| <i>PhD of Computer Science</i>         | Sep. 2023 - Sep. 2026 (Expected) |
| • <b>Shanghai Jiao Tong University</b> | Shanghai, China                  |
| <i>Master of Computer Science</i>      | Sep. 2017 - Mar. 2020            |
| • <b>Shanghai Jiao Tong University</b> | Shanghai, China                  |
| <i>Bachelor of Engineering</i>         | Sep. 2013 - Jun. 2017            |

## WORK EXPERIENCE

- **2020 – 2022:** NLP Researcher (full-time), Deep Learning for Code Generation and Matching, Tencent.
- **2019 – 2020:** NLP Research Intern, Deep Learning for Code Generation and Matching, Tencent.
- **2018 - 2019:** NLP Research Intern, Deep Learning for NLP and Recommender System, Microsoft Research.

## RESEARCH INTERESTS

- Understanding and Diagnosing LLMs through Mechanistic Interpretability.
- Enhancing LLM Capabilities through Mechanistic Understanding (multi-hop reasoning, arithmetic reasoning, catastrophic forgetting, in-context learning, model merging, model editing).

## RESEARCH SUMMARY

- **Understanding and Diagnosing LLMs through Mechanistic Interpretability** 2023 - Present
  - *Understanding internal mechanisms to enable faithful diagnosis of errors, shortcuts, and hallucinations*
  - **Developing VQALens for multimodal LLM diagnosis:** Developed an interpretability system to diagnose errors, reasoning shortcuts, and hallucinations, enabling fine-grained auditing of model predictions. (**System Demo**)
  - **Designing neuron-level attribution methods:** Designed neuron-level attribution methods to identify neurons responsible for knowledge storage and reasoning, supporting model diagnosis, editing, and bias analysis. (**EMNLP 2024, Main**)
  - **Analyzing in-context learning mechanisms:** Analyzed how LLMs perform in-context learning by identifying specialized attention heads that extract label features and implement metric learning mechanisms. (**EMNLP 2024, Main**)
- **Enhancing LLM Capabilities through Mechanistic Understanding** 2023 - Present
  - *From understanding to improvement: enhancing LLM capabilities based on mechanism analysis*
  - **Improving latent multi-hop reasoning in LLMs:** Improved latent multi-hop reasoning in LLMs through mechanism-guided post-training, including the proposed *Back Attention* module. (**EMNLP 2025, Main**)
  - **Model pruning for arithmetic reasoning:** Analyzed and pruned FFN neurons for arithmetic tasks based on mechanistic analysis and parameter identification. (**EMNLP 2024, Main**)
  - **Model merging for mitigating catastrophic forgetting:** Mitigated catastrophic forgetting in multimodal LLMs via neuron-level model merging after visual instruction tuning and RLHF. (**EMNLP 2025, Findings**)
  - **Model editing for reducing gender bias:** Reduced gender bias in LLMs through interpretable neuron-level model editing without degrading language performance. (**Under Review**)
- **Earlier Research Experience in Deep Learning** Before 2023
  - *Foundational work prior to my focus on LLMs: deep learning for code, NLP, recommender systems, and applied ML*
  - **Cross-modal retrieval for code representation learning:** Proposed cross-modal representation learning methods for binary source code matching by integrating CNNs and graph neural networks, resulting in a NeurIPS 2020 paper with 100+ citations and real-world deployment in industry systems. (**NeurIPS 2020**)
  - **Graph neural networks for binary code similarity detection:** Developed graph neural network models for binary code similarity detection, among the earliest deep learning approaches in this area, leading to an AAAI 2020 paper with 300+ citations and substantial follow-up work. (**AAAI 2020**)
  - **Adaptive user modeling for recommender systems:** Designed adaptive user modeling frameworks integrating long-term and short-term preferences, published at IJCAI 2019 and adopted in large-scale production recommender systems, which was used in MSN Advertising System. (**IJCAI 2019**)
  - **Efficient neural architectures for NLP:** Proposed sliced recurrent neural networks to improve training efficiency and scalability in NLP models, published at COLING 2018 with strong open-source impact. (**COLING 2018**)

## PUBLICATIONS AND PREPRINTS

---

- Locate-then-Merge: Neuron-Level Parameter Fusion for Mitigating Catastrophic Forgetting. **EMNLP 2025 Findings**  
**Zeping Yu**, Sophia Ananiadou.
- Back Attention: Understanding and Enhancing Multi-Hop Reasoning in Large Language Models. **EMNLP 2025 Main**  
**Zeping Yu**, Yonatan Belinkov, Sophia Ananiadou.
- Understanding and Mitigating Gender Bias in LLMs via Interpretable Neuron Editing. **Preprint**  
**Zeping Yu**, Sophia Ananiadou.
- Understanding Multimodal LLMs: The Mechanistic Interpretability of LLaVA in Visual Question Answering. **Preprint**  
**Zeping Yu**, Sophia Ananiadou.
- Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis. **EMNLP 2024 Main**  
**Zeping Yu**, Sophia Ananiadou.
- How do Large Language Models Learn In-Context? Query and Key Matrices of In-Context Heads are Two Towers for Metric Learning. **EMNLP 2024 Main**  
**Zeping Yu**, Sophia Ananiadou.
- Neuron-Level Knowledge Attribution in Large Language Models. **EMNLP 2024 Main**  
**Zeping Yu**, Sophia Ananiadou.
- CodeCMR: Cross-modal Retrieval for Function-Level Binary Source Code Matching. **NeurIPS 2020**  
**Zeping Yu**, Wenxin Zheng, Jiaqi Wang, Qiyi Tang, Sen Nie, Shi Wu.
- Order Matters: Semantic-Aware Neural Networks for Binary Code Similarity Detection. **AAAI 2020**  
**Zeping Yu\***, Rui Cao\*, Qiyi Tang, Sen Nie, Junzhou Huang, Shi Wu.
- Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation. **IJCAI 2019**  
**Zeping Yu**, Jianxun Lian, Ahmad Mahmoodi, Gongshen Liu, Xing Xie.
- Sliced Recurrent Neural Networks. **COLING 2018**  
**Zeping Yu**, Gongshen Liu.

## TECHNICAL SKILLS

---

- **Programming & Frameworks:** Python, PyTorch, TensorFlow, Keras, HuggingFace.
- **LLMs & Training:** Pretraining, SFT, RLHF, PEFT (LoRA), Model Editing, Model Merging.
- **Interpretability & Analysis:** Mechanistic Interpretability, Neuron Attribution, Token Attribution, Logit Analysis, Causal Analysis.
- **Models & Methods:** Transformers, GNNs, RNNs, Multimodal Models.