Zeping Yu

Homepage: https://zepingyu0512.github.io
Email: zepingyu@foxmail.com
Google Scholar: https://scholar.google.com/citations?user=OdpmpDsAAAAJ
Mobile: +44-7529930611
Target Position: Research Scientist / Applied Scientist (Large Language Models)
Wechat: +86-18621601510

Education

- **University of Manchester** Sep. 2023 - Sep. 2026
  PhD of Computer Science
- **Shanghai Jiao Tong University** Sep. 2017 - Mar. 2020
  Master of Computer Science
- **Shanghai Jiao Tong University** Sep. 2013 - Jun. 2017
  Bachelor of Engineering

Research Focus

- Mechanistic interpretability of large language and multimodal models, focusing on neuron-level analysis to guide model editing and merging for diagnosing and improving reasoning, reliability, and bias.
- Improving capabilities of LLMs and multimodal LLMs, including reasoning (factual knowledge, arithmetic, latent multi-hop reasoning), in-context learning, visual question answering, and robustness to catastrophic forgetting.

Research Experience

- **Analyzing and Diagnosing LLMs through Mechanistic Interpretability** 2023 - Present
  Mechanism-level analysis of LLM internals to diagnose hallucinations, shortcuts, and systematic errors
  - **VQALens: multimodal LLM diagnosis toolkit**: Built VQALens, a system for diagnosing multimodal LLM failures, which pinpoints error sources and reasoning shortcuts in VQA, enabling fine-grained auditing for LLMs. (**System Demo**)
  - **Neuron-level attribution for knowledge and reasoning**: Localized where factual knowledge is stored in LLMs by neuron-level attribution methods, enabling precise diagnosis and targeted editing for knowledge. (**EMNLP 2024, Main**)
  - **Mechanistic analysis of in-context learning**: Explained how in-context learning emerges in LLMs by identifying label-extracting attention heads that implement metric-learning–like behavior. (**EMNLP 2024, Main**)
- **Improving Capabilities of LLMs and Multimodal LLMs** 2023 - Present
  Mechanism-guided improvements for reasoning, robustness, and bias mitigation
  - **Improving latent multi-hop reasoning**: Improved multi-hop reasoning in LLMs by introducing a mechanism-guided Back Attention module that enhances multi-step information propagation during post-training. (**EMNLP 2025, Main**)
  - **Model merging for reducing catastrophic forgetting**: Mitigated catastrophic forgetting in multimodal LLMs via neuron-level parameter fusion after visual instruction tuning and RLHF. (**EMNLP 2025, Findings**)
  - **Model pruning for arithmetic reasoning**: Improved arithmetic reasoning while reducing model size by mechanistically identifying arithmetic-relevant FFN neurons and pruning redundant parameters. (**EMNLP 2024, Main**)
  - **Model editing for gender bias reduction**: Reduced gender bias via interpretable neuron-level model editing, maintaining core language performance while improving fairness-related behaviors. (**Under Review**)
- **Deep Learning for Code, NLP, and Recommender Systems** Before 2023
  Peer-reviewed research contributions across code intelligence, NLP, and recommender systems
  - **Cross-modal retrieval for function-level binary code matching**: Proposed cross-modal representation learning for function-level binary code matching by integrating neural and graph-based representations. (**NeurIPS 2020**)
  - **Graph neural models for binary code similarity detection**: Developed graph neural approaches for binary code similarity detection by modeling program structure and semantics. (**AAAI 2020**)
  - **Adaptive user modeling for personalized recommendation**: Designed adaptive user representation frameworks integrating long- and short-term preferences for personalized recommendation. (**IJCAI 2019**)
  - **Efficient neural architectures for NLP**: Proposed sliced recurrent neural networks to improve training efficiency and scalability for sequence modeling. (**COLING 2018**)

Work Experience

- **Deep Learning Research Engineer (Research-focused), Tencent** 2020 – 2022
  Industrial research on deep learning systems for code retrieval and understanding
  - **Model Design under Industrial Constraints**: Led the design and optimization of deep learning systems for function-level binary code retrieval, integrating CNNs and graph neural networks to model structured code representations at scale.
  - **Impact and Research Leadership**: As the primary contributor and first author, drove the project from problem formulation to large-scale evaluation, improving Recall@1 from 35.8% to 95.1%, achieving state-of-the-art performance and resulting in a **NeurIPS 2020** publication.

- **Deep Learning Research Intern, Tencent** 2019 – 2020
  Industrial research on graph-based neural models for code analysis
  - **Structured Model Design**: Developed semantic-aware graph neural network models for binary code similarity detection, explicitly modeling control-flow structure and instruction-level semantics.
  - **Impact and Research Ownership**: Led model development and evaluation as first author, improving Top-1 accuracy from 50% to 90% through systematic experimentation and ablation, and publishing the work at **AAAI 2020**.
- **Recommender System Research Intern, Microsoft Research Asia** 2018 – 2019
  Applied research on adaptive user modeling in large-scale recommender systems
  - **Model Design for User Representation**: Proposed adaptive user modeling frameworks that jointly capture long-term and short-term user preferences via temporal and content-aware controllers, extending RNN for large-scale recommendation.
  - **Impact and End-to-End Ownership**: Owned the project end-to-end as first author, validating the approach through both offline evaluation and large-scale online A/B testing, achieving 5–10% revenue gains and resulting in **IJCAI 2019**.

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 LLMs. **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: Mechanistic Interpretability of LLaVA in Visual Question Answering. **Preprint**
  Zeping Yu, Sophia Ananiadou.
- Interpreting Arithmetic Mechanism in LLMs 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 Mahmoody, 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, Logit Analysis, Causal Analysis.
- **Models & Methods**: Transformers, GNNs, RNNs, Multimodal Models.