Zijun Yao 姚子俊

yaozj20@mails.tsinghua.edu.cn +86 136 2565 5306 Google Scholar

Education

Ph.D. Tsinghua University **Sep. 2023 – Jun. 2026 (Expected)**
Knowledge Engineering Group, Department of Computer Science and Technology *Advised by Prof. Juanzi Li*

Master of Engineering. Tsinghua University **Sep. 2020 – Jun. 2023**
Knowledge Engineering Group, Department of Computer Science and Technology *Advised by Prof. Juanzi Li*

Bachelor of Engineering. Beijing University of Posts and Telecommunications **Sep. 2016 – Jun. 2020**
Major in Computer Science and Technology *Rank 1/321*

Work Experience

Zhipu.AI **Sep. 2021 – Now**
Research Topic: Building Large Language Models *Research intern*

NExT++ Research Centre, National University of Singapore **Mar. 2025 – Sep. 2025**
Research Topic: Trustworthy Large Language Models *Visiting scholar hosted by Prof. Tat-Seng Chua*

Qiyuan Lab **Mar. 2022 – Sep. 2022**
Research Topic: Knowledge Graph Construction *Research intern*

Knowledge Engineering Group, Tsinghua University **Jun. 2019 – Jun. 2020**
Research Topic: Graph Convolutional Networks *Research intern advised by Prof. Jie Tang*

Beijing University of Posts and Telecommunications **Mar. 2018 – Jun. 2020**
Research Topic: Evolutionary Game Theory on Graph *Research intern advised by Prof. Bin Wu*

Awards and Honors

ACL Best Demo Award **2023**
Awarded at ACL for outstanding system demonstration. First author paper. (***Top 1 / 155***)

CIKM Outstanding Resource Paper Nomination **2021**
Nominated for best resource paper at CIKM conference. (***Top 3 in Submissions***)

National Scholarship for Ph.Ds **2025**
Among Ph.D students in department of computer science and technology in THU. (***Top 15***)

National Scholarship for Undergraduates $(\times 3)$ **2017, 2018, 2019**
Awarded annually to top-performing undergraduate students (***Top 1%***)

Outstanding Graduate of Beijing **2020**
Honor given to top graduates across universities in Beijing

Tsinghua University Scholarships $(\times 4)$ **2024, 20233, 2022 2021**
University-level merit-based scholarships

First Prize in The Future Cup Activists of Artificial Intelligence & Robotic Projects **2021**
Competition award

Meritorious Winner of Mathematical Contest in Modeling **2018**
Competition award

Research

I am dedicated to exploring how to (1) establish the science of large language models (Science of LLMs); and (2) enable LLMs with knowledge and reasoning skills to solve scientific tasks (LLMs for Science).

Science of LLMs. While the powerful reasoning capabilities of LLMs are often attributed to emergent behaviors, their internal working mechanisms remain underexplored. To uncover the black-box nature of LLMs, I aim to establish a systematic science of LLMs. Microscopically, I investigate the connections between the fundamental building blocks of LLMs, such as hidden states, neurons, and sparse features—and their corresponding reasoning behaviors [6, 7]. At the macroscopic level, I study how the distinctive features of LLMs relate to key factors in model architecture design and the training process [8].

LLMs for Science. Scientific research seeks to advance the frontiers of human knowledge by addressing open research questions. With the rapidly evolving fact-seeking capabilities of LLMs, I aim to explore how to empower LLM agents to automate the scientific research process. My overall research plan consists of three components: (1) First, I investigate how to enhance the fact-seeking and question-answering capabilities of LLMs using both parametric and externally retrieved knowledge [1, 2, 3, 4]. (2) Second, I aim to develop an LLM-powered scientific research assistant system equipped with a comprehensive research environment, including a multi-agent framework and a set of evaluation protocol. (3) Finally, my long-term goal is to enable LLMs to autonomously improve their own research skills in real-world environments through interaction and feedback [5]. I am also exploring how to enable LLMs to investigate the science of LLMs themselves.

• Selected Publications

[1] **SeaKR: Self-aware Knowledge Retrieval for Adaptive Retrieval Augmented Generation**
Zijun Yao*, Weijian Qi*, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, Juanzi Li. In *ACL*, 2025. ***Oral Presentation (2.9% in Submission)***

[2] **VisKoP: Visual Knowledge oriented Programming for Interactive Knowledge Base Question Answering**
Zijun Yao*, Yuanyong Chen*, Xin Lv, Shulin Cao, Amy Xin, Jifan Yu, Hailong Jin, Jianjun Xu, Peng Zhang, Lei Hou, Juanzi Li. In *Demo of ACL*, 2023. ***Best Demo Award***

[3] **KoRC: Knowledge oriented Reading Comprehension Benchmark for Deep Text Understanding**
Zijun Yao*, Yantao Liu*, Xin Lv, Shulin Cao, Jifan Yu, Lei Hou, Juanzi Li. In *Findings of ACL*, 2023.

[4] **Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models**
Yantao Liu*, ***Zijun Yao****, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, Juanzi Li. In *LREC-COLING*, 2024.

[5] **RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style**
Yantao Liu, ***Zijun Yao***, Rui Min, Yixin Cao, Lei Hou, Juanzi Li. In *ICLR*, 2025. ***Oral Presentation (1.2% in Submission)***

[6] **Transferable and Efficient Non-Factual Content Detection via Probe Training with Offline Consistency Checking**
Xiaokang Zhang*, ***Zijun Yao****, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, Jie Tang. In *ACL*, 2024.

*Equal contribution.

[7] **LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder**
Yi Jing, ***Zijun Yao***, Lingxu Ran, Hongzhu Guo, Xiaozhi Wang, Lei Hou, Juanzi Li. In *EMNLP*, 2025.

[8] **How does Transformer Learn Implicit Reasoning?**
Jiaran Ye*, ***Zijun Yao***, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, Juanzi Li. In *NeurIPS* 2025. ***Spotlight (3.5% in Submission)***

[9] **Interpretable and Low-Resource Entity Matching via Decoupling Feature Learning from Decision Making**
Zijun Yao, Chengjiang Li, Tiansi Dong, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Yichi Zhang, Zelin Dai. In *ACL-IJCNLP*, 2021.

• Other Publications

[10] **LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking**
Amy Xin*, Yunjia Qi*, ***Zijun Yao***, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, Juanzi Li. In *CIKM*, 2025.

[11] **A General Neural-symbolic Architecture for Knowledge-intensive Complex Reasoning**
Shulin Cao*, ***Zijun Yao***, Lei Hou, Juanzi Li. In *Neurosymbolic Artificial Intelligence*, 2024.

[12] **FFAEval: Evaluating Dialogue System via Free-For-All Ranking**
Zeyao Ma*, ***Zijun Yao***, Jing Zhang, Jifan Yu, Xiaohan Zhang, Juanzi Li, Jie Tang. In Findings of EMNLP, 2024.

[13] **Dependency Parsing via Sequence Generation**
Boda Lin*, ***Zijun Yao***, Jiaxin Shi, Shulin Cao, Binghao Tang, Si Li, Yong Luo, Juanzi Li, Lei Hou. In *Findings of EMNLP*, 2022.

[14] **AtomR: Atomic Operator-empowered Large Language Models for Heterogeneous Knowledge Reasoning**
Amy Xin*, Jinxin Liu*, ***Zijun Yao***, Zhicheng Lee, Shulin Cao, Lei Hou, Juanzi Li. In *SIGKDD*, 2025.

[15] **SoAy: A Solution-based LLM API-using Methodology for Academic Information Seeking**
Yuanchun Wang, Jifan Yu, ***Zijun Yao***, Jing Zhang, Yuyang Xie, Shangqing Tu, Yiyang Fu, Youhe Feng, Jinkai Zhang, Jingyao Zhang, Bowen Huang, Yuanyao Li, Huihui Yuan, Lei Hou, Juanzi Li, Jie Tang. In *SIGKDD*, 2025.

[16] **Steering LVLMs via Sparse Autoencoder for Hallucination Mitigation**
Zhenglin Hua, Jinghan He, ***Zijun Yao***, Tianxu Han, Haiyun Guo, Yuheng Jia, Junfeng Fang. In *Findings of EMNLP*, 2025.

[17] **Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems**
Hao Peng, Yunjia Qi, Xiaozhi Wang, ***Zijun Yao***, Bin Xu, Lei Hou, Juanzi Li. In *ACL*, 2025.

[18] **Pre-training Distillation for Large Language Models: A Design Space Exploration**
Hao Peng, Xin Lv, Yushi Bai, ***Zijun Yao***, Jiajie Zhang, Lei Hou, Juanzi Li. In *ACL*, 2025.

[19] **Advancing Language Model Reasoning through Reinforcement Learning and Inference Scaling**
Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, ***Zijun Yao***, Juanzi Li, Jie Tang, Yuxiao Dong. In *ICML*, 2025.

[20] **KoLA: Carefully Benchmarking World Knowledge of Large Language Models**
Jifan Yu*, Xiaozhi Wang*, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, ***Zijun Yao***, . . ., Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, Juanzi Li. In *ICLR*, 2024.

[21] **DiaKoP: Dialogue-based Knowledge-oriented Programming for Neural-symbolic Knowledge Base Question Answering**
Zhicheng Lee, Zhidian Huang, ***Zijun Yao***, Jinxin Liu, Amy Xin, Lei Hou, Juanzi Li. In *Demo of CIKM*, 2024.

[22] **Evaluating Generative Language Models in Information Extraction as Subjective Question Correction**
Yuchen Fan*, Yantao Liu*, ***Zijun Yao***, Jifan Yu, Lei Hou, Juanzi Li. In *LREC-COLING*, 2024.

[23] **A Cause-Effect Look at Alleviating Hallucination of Knowledge-grounded Dialogue Generation**
Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, ***Zijun Yao***, Jing Zhang, Lei Hou, Juanzi Li. In *LREC-COLING*, 2024.

[24] **Probabilistic Tree-of-thought Reasoning for Answering Knowledge-intensive Complex Questions**
Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, ***Zijun Yao***, Qi Tian, Juanzi Li, Lei Hou. In *Findings of EMNLP*, 2023.

[25] **AKE-GNN: Effective Graph Learning with Adaptive Knowledge Exchange**
Liang Zeng*, Jin Xu*, ***Zijun Yao***, Yanqiao Zhu, Jian Li. In *CIKM*, 2023.

[26] **LittleMu: Deploying an Online Virtual Teaching Assistant via Heterogeneous Sources Integration and Chain of Teach Prompts**
Shangqing Tu, Zheyuan Zhang, Jifan Yu, Chunyang Li, Siyu Zhang, ***Zijun Yao***, Lei Hou, Juanzi Li. In *CIKM*, 2023.

[27] **GLM-dialog: Noise-tolerant Pre-training for Knowledge-grounded Dialogue Generation**
Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, ***Zijun Yao***, Zeyao Ma, Yiqi Xu, Haohua Wang, Xiaohan Zhang, Nianyi Lin, Sunrui Lu, Juanzi Li, Jie Tang. In *SIGKDD*, 2023.

[28] **MOOCRadar: A Fine-grained and Multi-aspect Knowledge Repository for Improving Cognitive Student Modeling in MOOCs**
Jifan Yu, Mengying Lu, Qingyang Zhong, ***Zijun Yao***, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, Juanzi Li, Jie Tang. In *SIGIR*, 2023.

[29] **Program Transfer for Answering Complex Questions over Knowledge Bases**
Shulin Cao, Jiaxin Shi, ***Zijun Yao***, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Zhiyuan Liu, Jinghui Xiao. In *ACL*, 2022.

[30] **MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs**
Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wenzheng Feng, Wei Xu, Shulin Cao, Kaisheng Zeng, ***Zijun Yao***, Lei Hou, Yankai Lin, Peng Li, Jie Zhou, Bin Xu, Juanzi Li, Jie Tang, Maosong Sun. In *CIKM*, 2021. ***Outstanding Resource Paper Nomination***

[31] **Calculating Biodiversity under Stochastic Evolutionary Dynamics**
Libin Zhang, ***Zijun Yao***, Bin Wu. In *Applied Mathematics and Computation*, 2021.

• Preprints

[32] **PairJudge RM: Perform Best-of-N Sampling with Knockout Tournament**
Yantao Liu, ***Zijun Yao***, Rui Min, Yixin Cao, Lei Hou, Juanzi Li. Submitted to *CoLM*, 2025. arXiv:2501.13007 (2025)

[33] **GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models**
GLM Team. arXiv:2508.06471 (2025)

[34] **Are Reasoning Models More Prone to Hallucination?**
Zijun Yao*, Yantao Liu*, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, Tat-Seng Chua. Submitted to *ARR Oct*, 2025. arXiv:2505.23646 (2025)

[35] **When Experimental Economics Meets Large Language Models: Tactics with Evidence**
Shu Wang, ***Zijun Yao***, Shuhuai Zhang, Jianuo Gai, Tracy Xiao Liu, Songfa Zhong. arXiv:2505.21371 (2025)

[36] **Toward Generalizable Evaluation in the LLM Era: A Survey Beyond Benchmarks**
Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, ***Zijun Yao***, Xiaozhi Wang, Dan Huang, Wenxuan Zhang, Lifu Huang, Muhao Chen, Lei Hou, Qianru Sun, Xingjun Ma, Zuxuan Wu, Min-Yen Kan, David Lo, Qi Zhang, Heng Ji, Jing Jiang, Juanzi Li, Aixin Sun, Xuanjing Huang, Tat-Seng Chua, Yu-Gang Jiang. arXiv:2504.18838 (2025)

[37] **Aligning Teacher with Student Preferences for Tailored Training Data Generation**
Yantao Liu, Zhao Zhang, ***Zijun Yao***, Shulin Cao, Lei Hou, Juanzi Li. arXiv:2406.19227 (2024)

Patent

1. **知识密集型推理问答方法、装置、电子设备和存储介质**
李涓子，姚子俊，曹书林，侯磊. 2024.
2. **一种实体记录匹配方法及系统**
李涓子，姚子俊，吕鑫，曹书林，陈源涌，史佳欣，侯磊，张鹏，唐杰，许斌. 2023.
3. **问答方法、装置、电子设备及存储介质**
李涓子，曹书林，史佳欣，姚子俊，吕鑫，于济凡，侯磊，张鹏，唐杰，许斌. 2023.
4. **问答推理方法及装置**
李涓子，姚子俊，侯磊，张鹏，唐杰，许斌. 2022.

Project

1. **面向知识密集型问答的神经符号推理研究:** 国家自然科学基金面上项目，2025.1 – 2028.12，项目成员
2. **大语言模型知识的表征、学习、记忆和注入机制分析与验证:** 北京市自然科学基金-小米创新联合基金，2024.7 – 2026.6，项目成员
3. **工业对话场景下面向知识的神经符号自然语言交互工具:** 企业合作项目（西门子），2023.9 – 2026.9，项目骨干
4. **基于预训练模型和多知识融合的开放域深度问答技术研究:** 企业合作项目（华为云），2023.9 – 2024.12，项目骨干
5. **面向分析师的知识工程技术:** 国家重点研发计划课题，2022.7 – 2024.6，项目成员
6. **工业知识图谱上的知识挖掘与推理:** 企业合作项目（西门子），2021.8 – 2022.7，项目骨干
7. **融合表示学习和符号规则的商品知识补全系统:** 企业合作项目（阿里），2021.3 – 2022.2，项目骨干
8. **文本蕴含的常识获取与融合:** 国家自然科学基金青年科学项目，2021.1 – 2023.12，项目成员
9. **大规模常识库的构建、表征、推理方法及开放平台:** 企业合作项目（国强研究院），2019.12 – 2024.12，项目骨干
10. **语义链接的多语言多模态知识图谱构建:** 北京智源人工智能研究院，2019.12 – 2021.12，项目骨干

System

GLM-zero: Reasoning model of the GLMs model family. Using reinforcement learning, it incentivizes long-thinking capability from the foundation model. *I build the training infrastructure for scalable reinforcement learning from verifiable reward.*

OpenSAE: Open-sourced sparse auto-encoder for interpreting the working mechanism of large language models. It integrates pre-training infrastructure and intervention operations for SAEs. It also open-sources 32 pre-trained SAE checkpoints. *I lead the project and develop the training, intervention, and analysis functions.*

VisKoP: Visual knowledge oriented programming platform for interactive knowledge Base question answering. VisKoP consists of an LLM-based interactive knowledge query interface and an efficient query execution engine on knowledge graph. *I lead the project and develop the high performance KG-query engine.*

Invited and Conference Talk

1. **Opening the Black Box: Understanding the Mechanism of LLMs:** Invited Talk, CS Frontier Tutorial, Peking University, 2025.
2. **揭开大语言模型内部机理的黑盒：大模型机理可解释性相关研究:** Invited Talk, 科学与艺术 Colloquium, Xinya College, Tsinghua University, 2025.
3. **Introduction to Large Language Models:** Invited Talk, Artificial Intelligence & Cybersecurity Workshop, National University of Singapore, 2025.
4. **Reinforcement Learning for Large Language Model Post-Training in a Nutshell:** Invited Talk, National University of Singapore, 2025; University of Science and Technology of China, 2025.
5. **Towards an Independent Researcher:** Invited Talk, Institute of Automation Chinese Academy of Science, 2025.
6. **SeaKR: Self-aware Knowledge Retrieval for Adaptive Retrieval Augmented Generation:** Oral Presentation, ACL-IJCNLP 2025, Vienna, Austria.
7. **Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models:** Oral Presentation, LREC-COLING 2024, Turin, Italy.
8. **VisKoP: Visual Knowledge oriented Programming for Interactive Knowledge Base Question Answering:** Poster Presentation, ACL 2023, Online.
9. **Interpretable and Low-Resource Entity Matching via Decoupling Feature Learning from Decision Making:** Oral Presentation, ACL-IJCNLP 2021, Online.

Teaching

As a teaching assistant, I have contributed to curriculum development, conducted tutorials, and assisted in designing course projects. Additionally, I delivered several lectures in the courses where I served as a teaching assistant. I also have the privilege of working with several exceptionally talented students below.

• Course

Natural Language Processing and Text Mining. **2025**
Tsinghua University *Teaching Assistant*

Knowledge Engineering. **2022, 2023, 2024**
Tsinghua University *Teaching Assistant*
Compilers Principles. **2019**
Beijing University of Posts and Telecommunications *Teaching Assistant*

• Lecture

1. **表示学习与语言模型 (Representation Learning and Language Modeling):** Natural Language Processing and Text Mining, Tsinghua University, 2025.
2. **Explainable Knowledge Question Answering: A Tutorial:** Knowledge Engineering, Tsinghua University, 2024.
3. **State-of-the-Art of Knowledge Engineering:** Knowledge Engineering, Tsinghua University, 2023.
4. **Explainable Knowledge Question Answering: A Tutorial:** Knowledge Engineering, Tsinghua University, 2023.
5. **Knowledge-oriented Programming Language for Complex Question Answering:** Knowledge Engineering, Tsinghua University, 2022.

• Mentoring

Current Students

1. **Yi Jing** (THU, Undergrad.): Linguistic mechanism of LLM.
2. **Lingxu Ran** (THU, Undergrad.): Constructing sparse autoencoder.
3. **Jinwu Hu** (THU, Undergrad.): Constructing sparse autoencoder.
4. **Hongzhu Guo** (PKU, Undergrad.): Linguistic mechanism of LLM.
5. **Amy Xin** (THU, Ph.D student): Scientific research agent.
6. **Zhidian Huang** (THU, master student): Reasoning mechanism of LLM.
7. **Jiaran Ye** (THU, master student): Reasoning mechanism of LLM.
8. **Yanxu Chen** (BUPT, master student): Scientific research agent.

Former Students

1. **Yantao Liu** (BUPT 2021, CAS-ICT 2022 – 2024): Knowledge integrated reasoning with LLMs; Reward modeling for reinforcement learning. Now Qwen reseacher.
2. **Zhicheng Lee** (THU 2023 – 2025): Knowledge base question answering. Now Zhipu.AI researcher.
3. **Weijian Qi** (XJTU 2024): Retrieval-augmented Generation. Now OSU, master student.
4. **Zeyao Ma** (RUC 2023): Evaluating LLMs with ranking. Now RUC, master student.
5. **Xiaokang Zhang** (RUC 2023 – 2024): Factual error detection in LLMs. Now Deepseek researcher.
6. **Yuchen Fan** (BUPT 2023 – 2024): LLMs generation evaluation. Now SJTU, Ph.D student.
7. **Yuanyong Chen** (THU 2022): Knowledge-base Question Answering. Now THU, master student.

Professional Service

- **Conference PC Member / Reviewer:** ACL Rolling Review (ARR); ACL; EMNLP; NeurIPS; ICLR; AISTATS; AAAI; IJCAI; CIKM
- **Conference Session Chair:** Information Extraction, ACL 2025
- **Journal Reviewer:** Transactions on Knowledge and Data Engineering (TKDE); Journal of Economic Behavior and Organization; (JEBO) Machine Intelligence Research (MIR); IEEE Transactions on Systems, Man and Cybernetics: Systems (SMC)
- **Other Activities:** Organizing Artificial Intelligence & Cybersecurity Workshop