Zekun Wang

Ph.D. Candidate in Research Center for SCIR, Harbin Institute of Technology

☎ +86 13523856329 • ✉ zkwang@ir.hit.edu.cn • 🌐 kugwzk.github.io

Research Interests

Model/Data-Efficiency & Acceleration: Efficient architecture for Transformers or hybrid ones, Pruning, distillation, quantization .etc to reduce model size and speedup inference, or efficient training LLMs or MLLMs with a small cost (time or data).
Multi-modal Models and Applications: Large multi-modal models (comprehensive, generation, or unify the both), which support diverse tasks and can be applied as agents in digital or embodied environments.

EDUCATION

Ph.D. Student, Harbin Institute of Technology **Harbin, China**
Major: Computer Science *Sept. 2019 - present*

B.E., Harbin Institute of Technology **Harbin, China**
Major: Software Engineering *Sept. 2015 - June 2019*

PUBLICATIONS

* denotes equal contributions

- Improved Diffusion-based Generative Model with Better Adversarial Robustness.
  Zekun Wang*, Mingyang Yi*, Shuchen Xue, Zhenguo Li, Ming Liu, Bing Qin, Zhi-Ming Ma.
  In: *Thirteenth International Conference on Learning Representations (ICLR).* 2025.
- AgentTrek: Agent Trajectory Synthesis via Guiding Replay with Web Tutorials.
  Yiheng Xu, Dunjie Lu, Zhennan Shen, Junli Wang, **Zekun Wang**, Yuchen Mao, Caiming Xiong, Tao Yu.
  In: *Thirteenth International Conference on Learning Representations (ICLR)* ***Spotlight***. 2025.
- CFSP: An Efficient Structured Pruning Framework for LLMs with Coarse-to-Fine Activation Information.
  Yuxin Wang*, Minghua Ma*, **Zekun Wang***, Jingchang Chen, Huiming Fan, Liping Shan, Qing Yang, Dongliang Xu, Ming Liu, Bing Qin. In: *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025).* 2025.
- Demons in the Detail: On Implementing Load Balancing Loss for Training Specialized Mixture-of-Expert Models.
  Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, **Zekun Wang**, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, Junyang Lin.
  Preprint. 2025.
- CodeElo: Benchmarking Competition-level Code Generation of LLMs with Human-comparable Elo Ratings.
  Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, **Zekun Wang**, Jian Yang, Zeyu Cui, Yang Fan, Yichang Zhang, Binyuan Hui, Junyang Lin.
  Preprint. 2025.
- Exploring & exploiting high-order graph structure for sparse knowledge graph completion.
  Tao He, Ming Liu, Yixin Cao, **Zekun Wang**, Zihao Zheng, Zheng Chu, Bing Qin.
  In *Journal of Frontiers of Computer Science.* 2025.
- Aguvis: Unified Pure Vision Agents for Autonomous GUI Interaction.
  Yiheng Xu*, **Zekun Wang***, Junli Wang*, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu,

Caiming Xiong.
Preprint. 2024.

- Qwen2.5 Technical Report.
  Qwen Team.
  Technical Report. 2024.
- CogGPT: Unleashing the Power of Cognitive Dynamics on Large Language Models.
  Yaojia Lv, Haojie Pan, **Zekun Wang**, Jiafeng Liang, Yuanxing Liu, Ruiji Fu, Ming Liu, Zhongyuan Wang, Bing Qin.
  In: *Findings of the Association for Computational Linguistics: EMNLP.* 2024.
- Divide-and-Conquer Meets Consensus: Unleashing the Power of Functions in Code Generation.
  Jingchang Chen, Hongxuan Tang, Zheng Chu, Qianglong Chen, **Zekun Wang**, Ming Liu, Bing Qin.
  In: *Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)* ***Oral***. 2024.
- GUIDE: A Guideline-Guided Dataset for Instructional Video Comprehension.
  Jiafeng Liang, Shixin Jiang, **Zekun Wang**, Haojie Pan, Zerui Chen, Zheng Chu, Ming Liu, Bing Qin, Ruiji Fu, Zhongyuan Wang.
  In: *33nd International Joint Conference on Artificial Intelligence (IJCAI).* 2024.
- SmartTrim: Adaptive Tokens and Attention Pruning for Efficient Vision-Language Models.
  Zekun Wang*, Jingchang Chen*, Wangchunshu Zhou, Haichao Zhu, Jiafeng Liang, Liping Shan, Ming Liu, Dongliang Xu, Qing Yang, Bing Qin.
  In: *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING-LREC)* ***Oral***. 2024.
- OFA-Diffusion Compression: Compressing Diffusion Model in One-Shot Manner.
  Zekun Wang, Mingyang Yi, Ming Liu, Bing Qin, Zhenguo Li.
  In Submission. 2024.
- MTGER: Multi-view Temporal Graph Enhanced Temporal Reasoning over Time-Involved Document.
  Zheng Chu, **Zekun Wang**, Jiafeng Liang, Ming Liu, Bing Qin.
  In *Findings of the Association for Computational Linguistics: EMNLP.* 2023.
- GTR: A Grafting-Then-Reassembling Framework for Dynamic Scene Graph Generation.
  Jiafeng Liang, Yuxin Wang, **Zekun Wang**, Ming Liu, Ruiji Fu, Zhongyuan Wang, Bing Qin.
  In *32nd International Joint Conference on Artificial Intelligence (IJCAI).* 2023.
- TAGNet: A Tiny Answer-Guided Network for Conversational Question Generation.
  Zekun Wang, Haichao Zhu, Ming Liu, Bing Qin.
  In: *International Journal of Machine Learning and Cybernetics (IJMLC).* 2023.
- Distilled Dual-Encoder Model for Vision-Language Understanding.
  Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, Furu Wei.
  In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2022.
- Less Is More: Domain Adaptation with Lottery Ticket for Reading Comprehension
  Haichao Zhu, **Zekun Wang**, Heng Zhang, Ming Liu, Sendong Zhao, Bing Qin.
  In: *Findings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2021.
- Molweni: A Challenge Multiparty Dialogues-based Machine Reading Comprehension Dataset with Discourse Structure.
  Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, **Zekun Wang**, Wenqiang Lei, Ting Liu, Bing Qin.
  In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING).* 2020.

EXPERIENCE

Alibaba Qwen Team. **Beijing, China**
Research Intern *May. 2024 - Present*

- Worked on pre-training and model architectures.

Huawei Noah's ark Lab. **Beijing, China**

Research Intern *Jul. 2023 - Apr. 2024*

- Worked on improving the efficiency of diffusion models by model compression and step acceleration.

Microsoft Research Asia (MSRA), Natural Language Computing Group. **Beijing, China**

Research Intern *Sept. 2020 - Sept. 2021*

- Worked with Researcher Wenhui Wang and Dr. Furu Wei on large-scale pre-trained models and efficient methods (like knowledge distillation) in multimodality and natural language processing.

Joint Laboratory of HIT and iFLYTEK Research, Reading Comprehension Group **Beijing, China**

Research Intern *June 2019 - Aug. 2019*

- Explored the generalization of pre-trained models on different question answering datasets and got the 3rd place in MRQA 2019 shared task@EMNLP 2019.

PROJECTS

Huozi Chat-LLM *Mar. 2023 - May. 2023*

- Participate in instruction turning with LLM and reduce resource consumption of deployment by pruning or quantization.

HONORS & AWARDS

Silver Medal **Shanghai, China**

The ACM/ICPC Asia Contest EC-Final *2017*

Silver Medal **Shenyang, China**

The ACM-ICPC Asia Regional Contest Shenyang Site *2017*

Silver Medal **Fuzhou, China**

The CCF Collegiate Computer Systems & Programming contest *2017*

National Scholarship **Harbin, China**

Harbin Institute of Technology *2018*

SERVICES

Reviewer: ICML, ICLR, NeurIPS, ACL, EMNLP, NAACL, ACL Rolling Review, COLING, WSDM, AAAI, IJCAI

SKILLS

Programming Languages	Python, C/C++
Tools and Frameworks	PyTorch, FairSeq, Huggingface Transformers, Pytorch Lightning