

## Education

### The Hong Kong Polytechnic University

*Ph.D. in Computer Science*

Advisor: Prof. Wenjie Li

Thesis Topic: Towards Lossless Inference Acceleration of Large Language Models

Jan. 2024 –

Hong Kong, China

### Peking University

*Master in Software Engineering*

Advisor: Prof. Zhifang Sui

Thesis: Speculative Decoding in Neural Machine Translation

Sep. 2020 – Jul. 2023

Beijing, China

### Peking University

*B.S. in Physics (Department of Astronomy)*

Advisor: Asst. Prof. Lijing Shao

Thesis: Improved Deep Learning Techniques in Gravitational-wave Data Analysis

Sep. 2016 – Jul. 2020

Beijing, China

## Preprints

- **Merlin's Whisper: Enabling Efficient Reasoning in LLMs via Black-box Adversarial Prompting**  
Heming Xia, Cunxiao Du, Rui Li, Chak Tou Leong, Yongqi Li, Wenjie Li
- **Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning**  
Xinghao Chen\*, Anhao Zhao\*, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, Xiaoyu Shen
- **LLM-REVal: Can We Trust LLM Reviewers Yet?**  
Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Junfeng liu, Xiangwen Kong, Zhifang Sui, and Nanyun Peng
- **HauntAttack: When Attack Follows Reasoning as a Shadow**  
Jingyuan Ma\*, Rui Li\*, Zheng Li, Junfeng Liu, Heming Xia, Lei Sha, and Zhifang Sui
- **From Query to Logic: Ontology-Driven Multi-Hop Reasoning in LLMs**  
Haonan Bian, Yutao Qi, Rui Yang, Yuanxi Che, Jiaqian Wang, Heming Xia, Ranran Zhen

## First-Author Publications

\* indicates equal contribution.

- **TokenSkip: Controllable Chain-of-Thought Compression in LLMs**  
Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, Wenjie Li  
*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. EMNLP 2025.*
- **Beyond Single Frames: Can LMMs Comprehend Temporal and Contextual Narratives in Image Sequences?**  
Xiaochen Wang\*, Heming Xia\*, Jialin Song, Longyu Guan, Yixin Yang, Qingxiu Dong, Weiyao Luo, Yifan Pu, Yiru Wang, Xiangdi Meng, Wenjie Li, Zhifang Sui  
*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. EMNLP 2025 (Findings).*
- **SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration**  
Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, Wenjie Li  
*The Thirteenth International Conference on Learning Representations. ICLR 2025.*
- **Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding**  
Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, Zhifang Sui  
*The 62nd Annual Meeting of the Association for Computational Linguistics. ACL 2024 (Findings).*
- **ImageNetVC: Zero- and Few-Shot Visual Commonsense Evaluation on 1000 ImageNet Categories**  
Heming Xia\*, Qingxiu Dong\*, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, Zhifang Sui  
*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. EMNLP 2023 (Findings).*
- **Bi-Drop: Enhancing Fine-tuning Generalization via Synchronous sub-net Estimation and Optimization**  
Shoujie Tong\*, Heming Xia\*, Damai Dai, Runxin Xu, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui  
*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. EMNLP 2023 (Findings).*

- **Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation**  
Heming Xia\*, Tao Ge\*, Peiyi Wang, Si-Qing Chen, Furu Wei, Zhifang Sui  
*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. EMNLP 2023 (Findings).*
- **Enhancing Continual Relation Extraction via Classifier Decomposition**  
Heming Xia, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui  
*The 61st Annual Meeting of the Association for Computational Linguistics. ACL 2023 (Findings, Short Paper).*
- **Improved deep learning techniques in gravitational-wave data analysis**  
Heming Xia, Lijing Shao, Junjie Zhao, Zhoujian Cao  
*Physical Review D 103 (2021), 024040.*

## Other Publications

---

- **KNN-SSD: Enabling Dynamic Self-Speculative Decoding via Nearest Neighbor Layer Set Optimization**  
Mingbo Song, Heming Xia, Jun Zhang, Chak Tou Leong, Qiancheng Xu, Wenjie Li, Sujian Li  
*The 19th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2026 (Findings).*
- **SpecVLM: Enhancing Speculative Decoding of Video LLMs via Verifier-Guided Token Pruning**  
Yicheng Ji\*, Jun Zhang\*, Heming Xia, Jinpeng Chen, Lidan Shou, Gang Chen, Huan Li  
*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. EMNLP 2025.*
- **Towards Harmonized Uncertainty Estimation for Large Language Models**  
Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, Zhifang Sui  
*The 63rd Annual Meeting of the Association for Computational Linguistics. ACL 2025 (Oral Presentation).*
- **How Far are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation**  
Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, Zhifang Sui  
*The 63rd Annual Meeting of the Association for Computational Linguistics. ACL 2025 (Findings).*
- **PEToolLLM: Towards Personalized Tool Learning in Large Language Models**  
Qiancheng Xu, Yongqi Li, Heming Xia, Fan Liu, Min Yang, Wenjie Li  
*The 63rd Annual Meeting of the Association for Computational Linguistics. ACL 2025 (Findings).*
- **AppBench: Planning of Multiple APIs from Various APPs for Complex User Instruction**  
Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, Kam-Fai Wong  
*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. EMNLP 2024.*
- **A Survey on In-context Learning**  
Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, Zhifang Sui  
*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. EMNLP 2024.*
- **Enhancing Tool Retrieval with Iterative Feedback from Large Language Models**  
Qiancheng Xu, Yongqi Li, Heming Xia, Wenjie Li  
*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. EMNLP 2024 (Findings).*
- **Taking a Deep Breath: Enhancing Language Modeling of Large Language Models with Sentinel Tokens**  
Weiyao Luo, Suncong Zheng, Heming Xia, Weikang Wang, Yan Lei, Tianyu Liu, Shuang Chen, Zhifang Sui  
*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. EMNLP 2024 (Findings).*
- **Can Large Multimodal Models Uncover Deep Semantics Behind Images?**  
Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, Zhifang Sui  
*The 62nd Annual Meeting of the Association for Computational Linguistics. ACL 2024 (Findings).*
- **Enhancing Continual Relation Extraction via Classifier Decomposition**  
Heming Xia, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui  
*The 61st Annual Meeting of the Association for Computational Linguistics. ACL 2023 (Findings, Short Paper).*
- **Lossless Acceleration for Seq2seq Generation with Aggressive Decoding**  
Tao Ge, Heming Xia\*, Xin Sun\*, Si-Qing Chen, Furu Wei  
*Microsoft Research Technical Report.*
- **Premise-based Multimodal Reasoning: Conditional Inference on Joint Textual and Visual Clues**  
Qingxiu Dong\*, Ziwei Qin\*, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, Zhifang Sui  
*The 60th Annual Meeting of the Association for Computational Linguistics. ACL 2022.*

# Internship

<b>Sea AI Lab</b> <i>Research Intern</i> Mentor: Dr. Cunxiao Du	<b>Jun. 2025 –</b> <i>Singapore</i>
<b>The Hong Kong Polytechnic University</b> <i>Research Assistant at NLP Group</i> Advisor: Prof. Wenjie Li	<b>Oct. 2023 – Jan. 2024</b> <i>Hong Kong, China</i>
<b>Microsoft Research Asia</b> <i>Research Intern at NLC Group</i> Mentor: Dr. Tao Ge	<b>Oct. 2021 – Aug. 2022</b> <i>Beijing, China</i>

## Services and Membership

- **Area Chair / Action Editor:** ACL, EMNLP, ACL ARR
- **Reviewer / Program Committee Member:** NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, ACM MM, EACL, AACL, TASLP, TWEB
- **Teaching Assistant:** COMP 5423: Natural Language Processing (Fall & Spring 2025), COMP 5140 (Fall 2024), COMP 2S01 (Spring 2024) at PolyU

## Open-Source Projects

- **Reading List for Speculative Decoding (1k Stars★):** Maintained a regularly updated paper list on Speculative Decoding, covering milestones, benchmarks, analytical studies, and applications on this promising research area.
- **Reading List for Efficient Reasoning (700 Stars★):** Maintained a curated paper list on efficient reasoning, covering efficient training, latent/long-to-short CoT, adaptive thinking, optimal test-time scaling strategies, and more.
- **Spec-Bench for Speculative Decoding (Python, PyTorch, 300 Stars★):** Developed a comprehensive benchmark and unified evaluation platform for assessing leading Speculative Decoding methods across diverse application scenarios.
- **Seq2Seq Inference Acceleration with Speculative Decoding (Python, Fairseq):** Released all the codes and checkpoints utilized in Speculative Decoding, which achieves 3x-5x inference speedup with only 300MiB of extra memory cost.
- **Deep Learning Toolkits for Gravitational-wave Analysis (Python, PyTorch):** Developed a deep learning toolkit for gravitational-wave (GW) data analysis, which supports GW data generation, visualization and classification.

## Invited Talks

- **NICE-NLP and MLNLP**, Sharing Panel - Efficient Reasoning in Large Language Models, 06/2025.
- **Huawei**, Hong Kong, Stop Overthinking: Towards Efficient Reasoning in Large Language Models, 05/2025.
- **COLING 2025 Tutorial**, Abu Dhabi, Speculative Decoding for Efficient LLM Inference, 01/2025.
- **CIP Group, CASIA**, Speculative Decoding: Past, Recent Advancements, and Future Directions, 04/2024.
- **NICE-NLP**, Unlocking Efficiency in LLM Inference: A Comprehensive Survey of Speculative Decoding, 03/2024.

## Technical Skills

<b>Languages:</b> Python, Latex, C/C++, Java, Shell, MATLAB, HTML/CSS
<b>Developer Tools:</b> PyCharm, VS Code, Git, Docker, Linux, Vim, Eclipse
<b>Libraries/Frameworks:</b> PyTorch, Transformers, Fairseq, TensorFlow, PyTorch-Lightning, spaCy, NumPy, WordPress

## Honors and Awards

• Merit Student, Peking University	2021
• Scholarship of National Astronomical Observatory, Chinese Academy of Sciences	2019
• Merit Student, Henan Province, China	2016