Zhepei (Bruce) Wei *Curriculum Vitae*

Contact Information

Department of Computer Science
University of Virginia
Room 224, Rice Hall
Charlottesville, VA 22903
Email : zhepei.wei@virginia.edu
Homepage: weizhepei.com

Research Interests

My recent research mainly focuses on large language model (LLM), including but not limited to retrieval-augmented generation (RAG), efficient inference, and alignment with human preference.

Education

University of Virginia — Charlottesville, U.S.
- Ph.D., Department of Computer Science — Aug 2022 - present
- Advisor: Yu Meng

Jilin University — Changchun, China
- M.S., School of Artificial Intelligence — Sep 2019 - Jun 2022
- B.S., College of Computer Science and Technology — Sep 2015 - Jul 2019
- Advisor: Yi Chang

Academic Experience

University of Virginia — Charlottesville, U.S.
Research Assistant — Aug 2023 - present
- **Advisor:** Yu Meng
- **Large Language Model (LLM):** Research on LLM alignment, retrieval-augmented generation, and efficient LLM inference.

University of Virginia — Charlottesville, U.S.
Research Assistant · Human-Centric Data Mining Lab — Aug 2022 - Aug 2023
- **Advisor:** Hongning Wang
- **Multi-agent Decision Making:** Research on incentivizing reinforcement learning (RL) agents for collaborative decision-making in federated environment.

Jilin University — Changchun, China
Research Assistant — Sep 2019 - Jun 2022
- **Advisor:** Yi Chang
- **Language Understanding and Reasoning:** Research on information extraction and natural language reasoning tasks with knowledge graph and pre-trained LMs.

Jilin University — Changchun, China
Research Assistant · Machine Learning Lab — May 2017 - Jun 2018
- **Advisor:** You Zhou
- **Autonomous Agent:** Research on improving deep Q-learning (reinforcement leaning) for autonomous decision-making in an interactive game environment.

Work Experience

Amazon — Seattle, U.S.
Applied Scientist Intern · Rufus Alignment Team — Jan 2025 - present
- **Manager:** Wenlin Yao, Lihong Li
- **Multi-modal Reasoning:** Work on improving the reasoning ability of multi-modal LLMs by scaling up inference-time compute.

Selected Publications

Google Scholar Citations: 700+, Github Stars: 900+ (as of Jan. 2025)

1. **Zhepei Wei**, Wei-Lin Chen, Xinyu Zhu, Yu Meng. *Fast and Accurate Language Model Decoding via Parallel Token Processing.* Under Review. Presented at NeurIPS 2024 AFM Workshop (Oral: 8/157).
2. Shangjian Yin, **Zhepei Wei**, Xinyu Zhu, Wei-Lin Chen, Yu Meng. *Self-Alignment Optimization for Language Models.* Under Review.
3. **Zhepei Wei**, Wei-Lin Chen, Yu Meng. *InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales.* **ICLR 2025**.
4. **Zhepei Wei**, Chuanhao Li, Tianze Ren, Haifeng Xu, Hongning Wang. *Incentivized Truthful Communication for Federated Bandits.* **ICLR 2024**.
5. **Zhepei Wei**, Chuanhao Li, Haifeng Xu, Hongning Wang. *Incentivized Communication for Federated Bandits.* **NeurIPS 2023**.
6. Erxin Yu, Lan Du, Yuan Jin, **Zhepei Wei**, Yi Chang. *Learning Semantic Textual Similarity via Topic-informed Discrete Latent Variables.* **EMNLP 2022**.
7. **Zhepei Wei**, Yue Wang, Jinnan Li, Zhining Liu, Erxin Yu, Yuan Tian, Xin Wang, Yi Chang. *Towards Unified Representations of Knowledge Graph and Expert Rules for Machine Learning and Reasoning.* **AACL-IJCNLP 2022**.
8. Runliang Niu, **Zhepei Wei**, Yan Wang, Qi Wang. *AttExplainer: Explain Transformer via Attention by Reinforcement Learning.* **IJCAI-ECAI 2022**.
9. **Zhepei Wei**, Yue Wang, Jinnan Li, Yingqi Guo, Zhining Liu, Erxin Yu, Yi Chang. *CogKG: Unifying Symbolic and Continuous Knowledge Representations for Machine Reasoning.* AKBC 2021 USKB workshop.
10. **Zhepei Wei**, Jianlin Su, Yue Wang, Yuan Tian, Yi Chang. *A Novel Cascade Binary Tagging Framework for Relational Triple Extraction.* **ACL 2020**.

Awards & Honors

- Copenhaver Charitable Trust Bicentennial Fellowship (2024) — UVA
- OpenAI Researcher Access Program Recipient (2024) — OpenAI
- John A. Stankovic Graduate Research Award (2024) — UVA
- NeurIPS Scholar Award (2023) — NeurIPS Foundation
- Computer Science Scholar Fellowship (2022) — University of Virginia (UVA)
- National Scholarship (2018, 2020) — Ministry of Education of China
- Outstanding Graduate Award (2020) — Jilin University
- Honor Graduation Award (2019) — Jilin University
- Outstanding Undergraduate Award (2016, 2017, 2018) — Jilin University
- National Endeavor Scholarship (2016) — Ministry of Education of China

Academic Services

Conference Paper Reviewer

- The International Conference on Learning Representations (ICLR), 2025
- The International Conference on Machine Learning (ICML), 2024, 2025
- Annual Conference on Neural Information Processing Systems (NeurIPS), 2024
- The AAAI Conference on Artificial Intelligence (AAAI), 2020, 2021, 2024, 2025
- The Web Conference (WWW), 2023
- Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021
- Annual Meeting of the Association for Computational Linguistics (ACL), 2020