Tao Sun

Github
Homepage
buaast@buaa.edu.com
(+86) 199-6723-0102

Education

Beihang University (BUAA) *Sep. 2023 - Jan. 2026 (expected)*

Master's Student in Computer Technology, School of Computer Science and Engineering Beijing, China

Supervisor: Zhoujun Li; Google Scholar Citation: 151 ; GPA: 86.32

Xiangtan University (XTU) *Sep. 2019 - Jun. 2023*

B.E. in Computer Science and Technology, School of Computer Science Xiangtan, Hunan, China

Rank: 2 / 88 ; GPA: 88.65

Selected Paper

BitsAI-CR: Automated Code Review via LLM in Practice. FSE 2025 Industry Track

- **Authors**: **Tao Sun**, Jian Xu, Yuanpeng Li, Zhao Yan, Ge Zhang, Lintao Xie, Lu Geng, Zheng Wang, Yueyan Chen, Qin Lin, Wenbo Duan, Kaixin Sui.
- Posted by Synced / 机器之心, a Top AI media in China. Invited by the CTO of SHEIN for an internal discussion.
- **Motivation**: Recognized the inefficiency and low precision of traditional and LLM-based **Code Review** processes in large-scale development, highlighting the need for a systematic and continuously improving solution.
- **Solution:** Designed a two-stage LLM pipeline leveraging a taxonomy of review rules and a data flywheel with Outdated Rate metrics, resulting in high-precision, continuously optimized code review adopted at scale in ByteDance.

UniCoder: Scaling Code Large Language Model via Universal Code. ACL 2024 Main Conference

- **Authors**: **Tao Sun***, Linzheng Chai*, Jian Yang*, Yuwei Yin, Hongcheng Guo, Jiaheng Liu, Bing Wang, Liqun Yang, Zhoujun Li.
- **Motivation**: Addressed limitations in **Code Generation** by introducing a language-agnostic universal code (UniCode) to bridge natural language and executable code.
- **Solution**: Demonstrated that using UniCode as an intermediate step with multi-task learning and and purpose-built 140K-sample dataset significantly boosts code generation performance.

P2P: Automated Paper-to-Poster Generation and Fine-Grained Benchmark ICLR 2025 Workshop

- **Authors**: **Tao Sun**, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongling Li, Ge Zhang, Wenhao Huang, Jian Yang, Zhoujun Li.
- Accepted to ICLR 2025 Workshop(non-archival). An extended version is under review at a top-tier conference.
- We present P2P, a multi-agent framework that leverages LLM-driven HTML **Code Generation** to transform research papers into polished posters, backed by a 30k-example instruction dataset and establish a fine-grained benchmark for rigorous evaluation.

Internship Experience

ByteDance Inc. - Seed - 字节跳动 Seed - Beijing, China *Sep. 2024 - Present*

- **Position**: Research Intern (Full-time, Onsite, Paid)
- Participated in development of the programming assistant for **Cici (豆包编程助手)**, China's leading LLM application.
- Spearheaded the development of code intent recognition systems and Code LLM for Cici (豆包).
- Involved in the construction of the internal Code Review system.

Meituan Inc. - 美团 - Beijing, China *Mar. 2024 - Aug.2024*

- **Position**: Research Intern (Full-time, Onsite, Paid)
- Conducted research and testing on Meituan's in-house Large Language Model, primarily focusing on enhancing the code expert model's capabilities in code repository-level completion and automated repair.
- Explored the application of LLM in the field of software engineering, including their abilities in handling long texts, code planning and testing.

Project

Research Projects in Progress

- **First Author**: *BitsAI-C^{2}R: Benchmarking, Automating, and Training Customized Code Review* (coming soon)

Open Source Contributions *Feb. 2018 - Present*

- **Collaborator**: ***Windrecorder(Github 3215 Stars)*** is a memory search app that records everything on your screen, to let you rewind what you have seen, query through OCR text or image description, and get activity statistics.
- **Owner**: ***myRime(Github 123 Stars)*** is a customized input method utilizing the Rime engine, suitable for use with Flypy Double Pinyin (Xiaohe Shuangpin), Luna Pinyin, iBus, Fcitx, Windows and MacOS.

Selected Honors and Awards

- China National Scholarship *2024*
- The First Prize Scholarship (Awarded by Beihang University) *2024*
- Pacemaker to Merit Student (Awarded by Xiangtan University, Top 2‰ in School) *2021*
- The Jingdong Scholarship (Awarded by Jingdong Inc.) *2022*
- The First Prize Scholarship (Awarded by Xiangtan University, Top 7% in School) *2020 & 2021 & 2022*
- Bronze Medal of 2021 ICPC Asia Regional Contest (Awarded by ICPC Foundation) *Nov. 2021*
- Silver Medal of 2021 CCPC National Invitational Contest (Awarded by Committee for CCPC) *Jun. 2021*
- Bronze Medal of 2021 CCPC Guilin Site Contest (Awarded by Committee for CCPC) *Nov. 2021*
- The First Prize of CUMCM in Hunan Division (Awarded by CSIAM) *Oct. 2021*

SERVICE

Teaching Assistant

- *Algorithm Training Team for ACM-ICPC*, School of Computer Science, Xiangtan University. *Summer 2020*

Reviewer

- ICONIP 2025, CIKM 2024, ICONIP 2024

Publications & Papers in Preparation

- **Tao Sun**, Jian Xu, Yuanpeng Li, Zhao Yan, Ge Zhang, Lintao Xie, Lu Geng, Zheng Wang, Yueyan Chen, Qin Lin, Wenbo Duan, Kaixin Sui. *BitsAI-CR: Automated Code Review via LLM in Practice.* arxiv: 2501.15134. **FSE 2025 Industry Track**.
- **Tao Sun***, Linzheng Chai*, Jian Yang*, Yuwei Yin, Hongcheng Guo, Jiaheng Liu, Bing Wang, Liqun Yang, Zhoujun Li. *UniCoder: Scaling Code Large Language Model via Universal Code.* **ACL 2024 Main Conference**.
- **Tao Sun**, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongling Li, Ge Zhang, Wenhao Huang, Jian Yang, Zhoujun Li. *P2P: Automated Paper-to-Poster Generation and Fine-Grained Benchmark.* arxiv: 2505.17104. **ICLR 2025 Workshop**. Under Review.
- **Tao Sun**, Dongsu Shen, Saiqin Long, Qingyong Deng, Shiguo Wang. *Neural Distinguishers on TinyJAMBU-128 and GIFT-64.* **ICONIP 2022 Oral**.
- **Tao Sun**, Yang Yang, Xianfu Cheng, Jian Yang, Yintong Huo, etc. *RepoFix: Real-World Repository-Level Code Evaluation: From Issue Detection to Bug Localization and Fixes.*
- Linzheng Chai*, Shukai Liu*, Jian Yang*, Yuwei Yin, Ke Jin, Jiaheng Liu, **Tao Sun**, Ge Zhang, Changyu Ren, Hongcheng Guo, Zekun Wang, Boyang Wang, Xianjie Wu, Bing Wang, Tongliang Li, Liqun Yang, Sufeng Duan, Zhoujun Li. *McEval: Massively Multilingual Code Evaluation.* arXiv 2406.07436. **ICLR 2025**.
- Linzheng Chai, Jian Yang, **Tao Sun**, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, Zhoujun Li. *xCOT: Cross-lingual Instruction Tuning for Cross-lingual Chain-of-Thought Reasoning.* arXiv: 2401.07037. **AAAI 2025**.
- Jiaheng Liu*, Zehao Ni*, Haoran Que*, **Tao Sun**, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, Zhaoxiang Zhang. *RoleAgent: Building, Interacting, and Benchmarking High-quality Role-Playing Agents from Script.* **NIPS 2024**.
- Xianfu Cheng*, Hang Zhang*, Jian Yang, Xiang Li, Weixiao Zhou, Kui Wu, Fei Liu, Wei Zhang, **Tao Sun**, Tongliang Li and Zhoujun Li. *XFormParser: Semi-structured Form Parser with multimodal and multilingual knowledge.* arXiv: 2405.17336. **COLING 2025**.
- Xianfu Cheng, Weixiao Zhou, Xiang Li, Jian Yang, Hang Zhang, **Tao Sun**, Wei Zhang, Yuying Mai, Tongliang Li, Xiaoming Chen and Zhoujun Li. *SVIPTR: Fast and Efficient Scene Text Recognition with Vision Permutable Extractor.* **CIKM 2024**.
- Shukai Liu*, Linzheng Chai*, Jian Yang*, Jiajun Shi, He Zhu, Liran Wang, Ke Jin, Wei Zhang, Hualei Zhu, Shuyue Guo, **Tao Sun**, Jiaheng Liu, Yunlong Duan, Yu Hao, Liqun Yang, Guanglin Niu, Ge Zhang, Zhoujun Li. *MdEval: Massively Multilingual Code Debugging.* arXiv 2411.02310. Under Review.