Yuxiang Wei

PhD Student
Department of Computer Science
University of Illinois at Urbana-Champaign

ywei40@illinois.edu
yuxiang.cs.illinois.edu
UniverseFly
(Last updated: Oct 22, 2024)

Research Interests

I am dedicated to building *code intelligence*, bridging software engineering and machine learning. My work focuses on designing advanced code models and programming systems to synthesize, repair, and test real-world software.

Notably, I lead the development of 🎩**Magicoder** [7] and **StarCoder2-Instruct** [4, 1], projects that have garnered **over 405k downloads and 2.1k GitHub stars**. The core techniques and datasets from these projects have been **adopted by leading industry language models**, including Meta's Llama 3.1, Google's CodeGemma, and IBM's Granite code models.

I also lead the **pretraining** of **Arctic-SnowCoder** [2], a high-performing small code model trained with progressively higher-quality data.

Education

Since 2022 **University of Illinois at Urbana-Champaign**, *Urbana, Illinois, USA*
PhD student in Computer Science. *Advisor: Prof. Lingming Zhang*
Anticipated graduation date: May, 2027

2017–2022 **Tongji University**, *Shanghai, China*
Bachelor of Engineering in Computer Science

Publications

[1] **Yuxiang Wei**, Federico Cassano, Jiawei Liu, Yifeng Ding, Naman Jain, Zachary Mueller, Harm de Vries, Leandro Von Werra, Arjun Guha, and Lingming Zhang. "Fully Transparent Self-Alignment for Code Generation". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: `https://openreview.net/forum?id=xXRnUU7xTL`. **NeurIPS'24**.

[2] **Yuxiang Wei**, Hojae Han, and Rajhans Samdani. *Arctic-SnowCoder: Demystifying High-Quality Data in Code Pretraining*. 2024. arXiv: `2409.02326 [cs.CL]`. URL: `https://arxiv.org/abs/2409.02326`.

[3] Jiawei Liu, Songrun Xie, Junhao Wang, **Yuxiang Wei**, Yifeng Ding, and LINGMING ZHANG. "Evaluating Language Models for Efficient Code Generation". In: *First Conference on Language Modeling*. 2024. URL: `https://openreview.net/forum?id=IBCBMeAhmC`. **COLM'24**.

[4] **Yuxiang Wei**, Federico Cassano, Jiawei Liu, Yifeng Ding, Naman Jain, Harm de Vries, Leandro von Werra, Arjun Guha, and Lingming Zhang. *StarCoder2-Instruct: Fully Transparent and Permissive Self-Alignment for Code Generation*. `https://huggingface.co/blog/sc2-instruct`. 2024.

[5] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, **Yuxiang Wei**, et al. "StarCoder 2 and The Stack v2: The Next Generation". In: *arXiv preprint arXiv:2402.19173* (2024).

[6] Yifeng Ding, Jiawei Liu, **Yuxiang Wei**, and Lingming Zhang. "XFT: Unlocking the Power of Code Instruction Tuning by Simply Merging Upcycled Mixture-of-Experts". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12941–12955. URL: https://aclanthology.org/2024.acl-long.699. **ACL'24**.

[7] **Yuxiang Wei**, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. "Magicoder: Empowering Code Generation with OSS-Instruct". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024, pp. 52632–52657. URL: https://proceedings.mlr.press/v235/wei24h.html. **ICML'24**.

[8] **Yuxiang Wei**, Chunqiu Steven Xia, and Lingming Zhang. "Copiloting the Copilots: Fusing Large Language Models with Completion Engines for Automated Program Repair". In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. San Francisco, CA, USA: Association for Computing Machinery, 2023, pp. 172–184. ISBN: 9798400703270. DOI: 10.1145/3611643.3616271. URL: https://doi.org/10.1145/3611643.3616271. **ESEC/FSE'23**.

[9] Chunqiu Steven Xia, **Yuxiang Wei**, and Lingming Zhang. "Automated Program Repair in the Era of Large Pre-Trained Language Models". In: *Proceedings of the 45th International Conference on Software Engineering*. ICSE '23. Melbourne, Victoria, Australia: IEEE Press, 2023, pp. 1482–1494. ISBN: 9781665457019. DOI: 10.1109/ICSE48619.2023.00129. URL: https://doi.org/10.1109/ICSE48619.2023.00129. **ICSE'23**.

[10] Jiawei Liu, **Yuxiang Wei**, Sen Yang, Yinlin Deng, and Lingming Zhang. "Coverage-Guided Tensor Compiler Fuzzing with Joint IR-Pass Mutation". In: *Proc. ACM Program. Lang.* 6.OOPSLA1 (Apr. 2022). DOI: 10.1145/3527317. URL: https://doi.org/10.1145/3527317. **OOPSLA'22**.

Experiences

Since 08/2024 **Meta AI, Code Llama Team**, *Research Scientist Intern (through Magnit)*
Building an agentic code model to solve real-world software engineering tasks.
Hosted by Sida Wang

05–08/2024 **Snowflake GenAI, Arctic Training Team**, *Research Intern*
Pretrained SnowCoder [2] by demystifying and leveraging high-quality data insights.
Hosted by Rajhans Samdani, Kelvin So, Yusuf Ozuysal, and Yuxiong He

Since 11/2023 **BigCode Project**, *Member*
Led StarCoder2-Instruct [4, 1] and contributed to StarCoder2 [5].

Since 08/2022 **University of Illinois at Urbana-Champaign**, *Research Assistant*
Developing code intelligence through the synergy of software engineering and machine learning.
Advised by Lingming Zhang

Academic Services

(OC: Organizing Committee, AEC: Artifact Evaluation Committee)

OC International Workshop on Large Language Models for Code (LLM4Code'25)
OC International Workshop on Large Language Models for Code (LLM4Code'24)
Reviewer International Conference on Learning Representations (ICLR'25)
Reviewer IEEE Transactions on Software Engineering (TSE)
Reviewer Annual Conference on Neural Information Processing Systems (NeurIPS'24)
Reviewer IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'24)
Reviewer Great Lakes Symposium on VLSI (GLSVLSI'24)
Reviewer Workshop on Synthetic Data for Computer Vision (SynData4CV@CVPR'24)
Reviewer Workshop on Reliable and Responsible Foundation Models (R2-FM@ICLR'24)

AEC ACM SIGSOFT Symposium on Software Testing and Analysis (ISSTA'24)
AEC ACM Conference on Computer and Communications Security (CCS'23)
AEC Programming Language Design and Implementation (PLDI'24)

Invited Talks

Sep 2024 **Guest Lecture on Language Models for Code**
CS6501 @ University of Virginia

Mar 2024 **Discussions on Magicoder and Its Extensions**
Meta AI (Code Llama)

Jan 2024 **Magicoder: Source Code Is All You Need**
Snowflake GenAI (Copilot)

Oct 2023 **Fusing Large Language Models with Completion Engines for Code Generation**
Kwai Inc.

Apr 2023 **Combining Large Language Models with Symbolic Methods**
Uber Programming Systems Lab

Selected Awards

Sep 2024 Selected Proposal, Amazon Trusted AI Challenge ($250,000)
Jun 2024 OpenAI Researcher Access Program ($5000)
Oct 2023 NSF Student Travel Award ($1800)
Oct 2023 ACM SIGSOFT CAPS Award ($400)
Mar 2021 1st Prize of "Challenge Cup" Academic Works Competition, Tongji University
Nov 2019 National 2nd Prize (3.84%) of Chinese Mathematical Contest in Modeling
Nov 2019 Province-Level 1st Prize (Shanghai) of Chinese Mathematical Contest in Modeling

Open-Source Contributions

I enjoy developing and sharing high-quality open-source tools:

- **StarCoder2-Instruct** [4, 1] [GitHub] (**210+ stars, 15k downloads**): the very first entirely self-aligned code LLM trained with a fully permissive and transparent pipeline, surpassing CodeLlama-70B-Instruct on HumanEval.
- 🎩**Magicoder** [7] [GitHub] (**2k stars, 393k downloads**): enhancing code generation with *OSS-Instruct*, surpassing ChatGPT on HumanEval+ with $\leq$ 7B parameters.
- **Repilot** [8] [GitHub] (**120+ stars**): patch/code generation by combining large language models and semantics-based completion engines.
- **Tzer** [10] [GitHub] (**70+ stars**): fuzzer for the low-level IR (Intermediate Representation) of the TVM machine learning compiler.