Yifan Zhang

U.S. Green Card holder | Homepage | LinkedIn | Twitter | Email | Phone: +1 919 638-0501

Research Interests

My research focuses on AI for Code and Data-Centric AI for Systems. I aim to develop reliable systems by integrating SFT and RLHF in LLMs with neural-symbolic code representations, leveraging both static and runtime data to enhance system security and optimization. I am also interested in AI applications, including databases, cyber-physical systems, and medical imaging.

Education

Vanderbilt University — *Nashville, TN, USA*
Ph.D. in Computer Science, Specialize in AI for Software Engineering — *Jun. 2022 – Dec. 2026 (Expected)*
- Advised by Prof. Kevin J. Leach & Prof. Yu Huang

Georgia Institute of Technology — *Atlanta, GA, USA*
M.Sc. in Computer Science, Specialize in Machine Learning — *Aug. 2022 – Jun. 2025*
- Studied courses in RL and program analysis

China University of Petroleum — *Beijing, CN*
M.Eng. in Petroleum Engineering, Specialize in Industrial & System Engineering — *Sep. 2012 – Jun. 2019*
- Double majored in English (TEM-8 holder) & minored in British Parliamentary debates

University of Calgary — *Calgary, AB, CA*
Undergraduate Exchange Program in Petroleum Engineering — *Dec. 2015 – Jun. 2016*
- Fully-funded by Chinese national fellowship for overseas studies

Publications (*Alphabetical; †Mentor)

Refereed Journal & Conference Papers

[R1] Jiliang (Eric) Li*, **Yifan Zhang***†, Yu Huang, Kevin Leach: *MalMixer: Few-Shot Malware Classification with Retrieval-Augmented Semi-Supervised Learning.* In Proceedings of the 10th IEEE European Symposium on Security and Privacy (**EuroS&P'25**), June 30-July 4, 2025. CORE A. To Appear. [Paper] [arXiv] [artifact]

[P1] **Yifan Zhang**, Chen Huang, Zachary Karas, Thuy Dung Nguyen, Kevin Leach, Yu Huang: *Enhancing Code LLM Training with Programmer Attention.* In Companion Proceedings of the ACM International Conference on the Foundations of Software Engineering (**FSE'25**). June 23-27, 2025. IVR Paper (Companion). To Appear. [Paper] [arXiv]

[P2] **Yifan Zhang**, Xue (Sharon) Yang: *CONSTRUCTA: Automating Commercial Construction Schedules in Fabrication Facilities with Large Language Models.* In Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (**NAACL'25**). April 29–May 4, 2025. Industry Paper. CORE A. To Appear. [Paper] [arXiv] [poster] [slides] [video]

[P3] Manish Acharya*, **Yifan Zhang***†, Kevin Leach, Yu Huang: *Optimizing Code Runtime Performance through Context-Aware Retrieval-Augmented Generation.* In Proceedings of the 33rd IEEE/ACM International Conference on Program Comprehension (**ICPC'25**). April 27-28, 2025. ERA Paper. CORE A. To Appear. [Paper] [arXiv] [artifact]

[P4] Yichang He, **Yifan Zhang**, Yunpeng Fan, U-Xuan Tan: *Real-Time Vibration Compensation with Long Short-Term Memory Recurrent Neural Network and Adaptive Filter.* The IEEE/ASME Transactions on Mechatronics (**TMech**), 2024. SJR Q1. [Paper].

[P5] Zachary Karas, Aakash Bansal, **Yifan Zhang**, Jia-Jun (Toby) Li, Collin McMillan, Yu Huang: *A Tale of Two Comprehensions? Studying Human Attention during Code Summarization.* The ACM Transactions on Software Engineering and Methodology (**TOSEM**), 2024. SJR Q1. **ICSE'25 Journal First Paper**. [Paper] [artifact]

[P6] **Yifan Zhang**, Jiliang (Eric) Li, Zachary Karas, Aakash Bansal, Jia-Jun (Toby) Li, Collin McMillan, Kevin Leach, Yu Huang: *EyeTrans: Merging Human and Machine Attention for Neural Code Summarization.* In

Proceedings of the ACM International Conference on the Foundations of Software Engineering (**FSE'24**), July 15-19, 2024. CORE A*. [Paper] [arXiv] [artifact] [poster] [slides]

[P7] Chen Huang, Haoyang Li, **Yifan Zhang**, Wenqiang Lei, Jiancheng Lv: *Cross-Space Adaptive Filter: Integrating Graph Topology and Node Attributes for Alleviating the Over-smoothing Problem.* In Proceedings of the ACM Web Conference (**WWW'24**), May 13-17, 2024. CORE A*. [Paper] [arXiv] [poster] [slides] [video]

[P8] Jiliang (Eric) Li, **Yifan Zhang**†, Zachary Karas, Collin McMillan, Kevin Leach, Yu Huang : *Do Machines and Humans Focus on Similar Code? Exploring Explainability of Large Language Models in Code Summarization.* In Proceedings of 32nd IEEE/ACM International Conference on Program Comprehension (**ICPC'24**), April 15-16, 2024. RENE Paper. CORE A. [Paper] [arXiv] [slides]

[P9] Haoyu Dong, **Yifan Zhang**, Hanxue Gu, Nicholas Konz, Maciej Mazurowski: *SWSSL: Sliding-Window based Self-Supervised Learning Framework for Anomaly Detection.* The IEEE Transactions on Medical Imaging (**TMI**), 2023. SJR Q1. [Paper] [artifact]

[P10] Aakash Bansal, Chia-Yi Su, Zachary Karas, **Yifan Zhang**, Yu Huang, Jia-Jun (Toby) Li, Collin McMillan: *Modeling Programmer Attention as Scanpath Prediction.* In Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering (**ASE'23**), September 11-15, 2023. NIER Paper. CORE A*. [Paper] [arXiv]

[P11] **Yifan Zhang**: *Leveraging Artificial Intelligence on Binary Code Comprehension.* In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (**ASE'22**), October 10-14, 2022. Doctoral Symposium Paper. CORE A*. [Paper] [arXiv] [poster] [slides]

[P12] **Yifan Zhang***, Haoyu Dong*, Nicholas Konz, Hanxue Gu, Maciej Mazurowski: *Lightweight Transformer Backbone for Medical Object Detection.* In Workshop Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention Workshops, September 22nd, 2022. (**MICCAIW'22**), 2022. [Paper] [arXiv] [poster] [slides] [video]

[P13] Jing Li, Xiangfang Li, Keliu Wu, Dong Feng, Tao Zhang, **Yifan Zhang**: *Thickness and Stability of Water Film Confined inside Nanoslits and Nanocapillaries of Shale and Clay.* The International Journal of Coal Geology (**IJCG**), 2017. SJR Q1. [Paper]

Papers In Submitting & Under Review

[R2] **Yifan Zhang**, Kevin Leach: *Human-Inspired LLMs for Neural Software Engineering.* Under review at the ACM International Conference on the Foundations of Software Engineering Workshops (FSEW'25). Position Paper.

[R3] **Yifan Zhang**, Kevin Leach: *Leveraging Human Insights for Enhanced LLM-based Code Repair.* Under review at the ACM International Conference on the Foundations of Software Engineering Workshops (FSEW'25).

[R4] **Yifan Zhang**, Chen Huang, Yueke Zhang, Kevin Cao, Scott Thomas Anderson, Huajie (Jay) Shao, Kevin Leach, Yu Huang: *Pre-Training Representations of Binary Code Using Contrastive Learning.* In submitting to Transactions on Machine Learning Research (TMLR), 2025. [arXiv] [artifact]

[R5] **Yifan Zhang**, Jiliang (Eric) Li, Kexin Pei, Yu Huang, Kevin Leach: *VulRAG: Code Vulnerability Repair by Retrieval-Augmented Generation.* In submitting to the 40th IEEE/ACM International Conference on Automated Software Engineering (ASE'25).

[R6] Suad Mohamed, Zachary Karas, **Yifan Zhang**†, Yu Huang: *The Potential of Human Attention Patterns to Predict Code Summary Quality.* In submitting to the 41st IEEE International Conference on Software Maintenance and Evolution (ICSME'25).

Non-Archival Workshops & Symposiums Papers

[N1] **Yifan Zhang***, Michael Sandborn*, Stefan Larson, Yu Huang, Kevin Leach: *K-ASTRO: Structure-Aware Adaptation of LLMs for Code Vulnerability Detection.* Presented at the Hot Topics in the Science of Security Symposium 2025 (HotSoS'25). April 2-4, 2025. WIP Paper. [webpage] [slides]

[N2] **Yifan Zhang**, Junwen Yang, Haoyu Dong, Qingchen Wang, Huajie (Jay) Shao, Kevin Leach, Yu Huang: *ASTRO: An AST-Assisted Approach for Generalizable Neural Clone Detection.* Presented at the 45th International Conference on Software Engineering Workshops (ICSEW'23), May 14th, 2023. Short Paper. [WIP] [slides]

[N3] **Yifan Zhang**, Haoyu Dong, Nicholas Konz, Hanxue Gu, Maciej Mazurowski: *REPLICA: Enhanced Feature Pyramid Network by Local Image Translation and Conjunct Attention for High-Resolution Breast Tumor Detection.* Updated to be in the 25th MICCAI Workshops. [WIP]

Industry Experience

IBM T.J. Watson Research *May. 2025 – Aug. 2025*

Incoming Research Intern, Supervisor: Dr. Elham Khabiri & Dr. Achille Fokoue *Yorktown Heights, NY, USA*

- **Knowledge LLM**: Planned project focusing on advancing AI reasoning capabilities. Aiming to design a novel cognitive framework that leverages large-scale language models for deep reasoning, context-aware decision making, and enhanced inference efficiency. Collaborating with interdisciplinary teams to explore cutting-edge AI methodologies in industrial applications.

Google LLC *Jan. 2025 – Apr. 2025*

Student Researcher, Supervisor: Dr. Mircea Trofin & Dr. Xinliang (David) Li *Sunnyvale, CA, USA*

- **LLVM IR Benchmark Automation**: Designed an automated pipeline using LLMs to generate C++ code that simulates LLVM IR profiles, enabling benchmarking and performance analysis of production code. The system supports AI models in learning compiler behavior and optimization strategies across diverse hardware platforms, with safeguards in place to preserve IP integrity.

Intel Corporation *Jun. 2024 – Dec. 2024*

GenAI Research Intern, Supervisor: Dr. Jun Li & Dr. Xue (Sharon) Yang *Santa Clara, CA, USA*

- **Construction Schedule Automation**: Developed and implemented a LLM-based system to automate and enhance construction scheduling for Intel's semiconductor fabrication projects. Integrated rule-based knowledge construction and in-context learning to predict and adjust schedules, leading to improved 2.8x efficiency compared with GPT-4o and reduced manual intervention.

TikTok/ByteDance *May. 2023 – Aug. 2023*

Research Scientist Intern, Supervisor: Dr. Zhibing Zhao & Dr. Tieying Zhang *San Jose, CA, USA*

- **SQL Hint Recommendation**: Enhancing the SQL Pipeline through a Hint Recommendation System Based on Representation Learning: The model leverages two Transformer models in a pipeline and pretrains a generalizable model for reranking SQL plans across diverse table spaces and schemas.
- **Research Collaboration**: Participated in group research activities, including giving technical talk at TikTok InfraLab and working with other team members in LLM literature review and conference paper review.

JD.com *Dec. 2018 – Mar. 2021*

Machine Learning Engineer, Supervisors: Dr. Hu Wang & Mr. Chen Huang *Beijing, CN*

- **Action Model**: Built Bi-GRU and DeepFM models on user behavior features to predict the credit use rate and overall profit of every user in cash loan and consumer debt. The model can propose decisions to increase their credit limit for maximizing income, and achieved 21.4% overall profit increase.
- **Credit Score Propagation**: Built a heterogeneous graph on different types of user connections, and applied GNN models to propagate the credit score and improve risk prediction. The model can improve the overall accuracy of the XGB model by 5% in user classification.
- **Privacy-Preserving Collaboration**: Invented one kind of GAN-styled model using differential privacy to improve the efficiency and security of federated learning. Granted 5 CN patents based on the research outputs. One of the patents was awarded as 1st Runner-up in the 3rd JD Discovery Cup Patent Competition.

Academic Experience

Duke University *Jul. 2021 – Jun. 2022*

Research Associate, Advisor: Prof. Maciej A. Mazurowski *Durham, NC, USA*

- **Medical Object Detection**: Designed a feature interpolation pipeline for injecting tumors into healthy images as an augmented dataset, and conjuncted a ViT on the outputs of a ResNet as inputs to a FPN in Faster R-CNN for tumor detection. The model mitigates the data-hungry problem of attention and achieves 13.1% improvement in AP50 for detecting tumors.

University of Hong Kong *Mar. 2021 – Sep. 2021*

Senior Research Assistant, Advisor: Prof. Qingchen Wang *Hong Kong SAR*

- **Financial Decision Making**: Built an entire intelligent debt collection system using data-driven deep reinforcement learning models. The model utilizes Transformer as the feature extractor and attaches a offline policy gradient model trained on the embedded sequential-aware hidden features to propose long-term dependent decisions.

Academic Services

Program Committee (PC) Member

- 2025 Annual Meeting of the Association for Computational Linguistics (ACL'25) Industry Track
- 2025 Annual Conference on Neural Information Processing Systems (NeurIPS'25)
- 2025 International Conference on Machine Learning (ICML'25)
- 2025 International Conference on Learning Representations (ICLR'25)
- 2024 Annual Conference on Neural Information Processing Systems (NeurIPS'24) Datasets & Benchmarks Track
- 2024 Annual Conference on Neural Information Processing Systems (NeurIPS'24) [webpage]
- 2024 ACM Conference on Computer and Communications Security (CCS'24) Artifact Evaluation Track [webpage]
- 2024 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'24)
- 2024 Data-Centric Machine Learning Research Workshop at ICLR (DMLR@ICLR'24) with **Exceptional Reviewer Award** (Top 9.43% venue-wide) [webpage]
- 2024 AAAI Conference on Artificial Intelligence (AAAI'24)
- 2023 Annual Conference on Machine Learning and Systems (MLSys'23) Artifact Evaluation Track
- 2023 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'23)
- 2023 AAAI Workshop on DL-Hardware Co-Design for AI Acceleration (DCAA@AAAI'23)
- 2023 International Conference on Mining Software Repositories (MSR'23)
- 2023 AAAI Conference on Artificial Intelligence (AAAI'23)
- 2022 MICCAI Workshop on Cancer Prevention through Early Detection (CaPTion@MICCAI'22)

External Conference Reviewer

- 2025 The annual IEEE International Conference on Data Engineering (ICDE'25) Industry Track
- 2025 International Conference on Extending Database Technology (EDBT'25) Industry Track
- 2025 Network and Distributed System Security (NDSS'25)
- 2024 International Conference on Very Large Data Bases (VLDB'24) Scalable Data Science Track
- 2023 USENIX Security Symposium (USENIX Security'23)
- 2022 IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR'22)

Research and Instructional Mentorship

Research Mentorship Experience

- Zichen (Ryan) Zhu, Bachelor of Applied Math & CS at Vanderbilt University. From 2025 Spring to Present. Research Topic: Dynamic Sparsity in Code LLM Training.
- Andrew B. Liu, Bachelor of Math & CS at Vanderbilt University. From 2025 Spring to Present. Research Topic: Reinforcement Learning for Intelligent Programming Guidance.
- Suad Mohamed, Bachelor of Psychology & CS at Belmont University. *Former Software Engineer Intern at Microsoft.* Co-Advised with Mr. Zachary Karas. From 2024 Spring to Present. Research Topic: Eye-Tracking for Human-Centered AI in Code Summarization. First Placement: Software Engineer at Microsoft.
- Manish Acharya, Bachelor of Math & CS at Vanderbilt University. *Vanderbilt Chancellor's Scholar.* From 2024 Spring to Present. Research Topic: Software Patch In-Context Learning.
- Jieyu Li, Bachelor of EEE at Shanghai Jiaotong University (SJTU), Master student of ECE at Vanderbilt University. From 2023 Fall to Present. Research Topic: AI Agent for Automated Program Repair. First Placement: Ph.D. in ECE at University of Waterloo. Advisor: Prof. Weiyi (Ian) Shang.
- Luka Mushkudiani, Bachelor of Math & CS at Vanderbilt University. *Former Intern at Meta's PyTorch Compiler Team.* From 2023 Fall to Present. Research Topic: Neural Type Systems for Python/Javascript
- Jiliang (Eric) Li, Bachelor of Math & CS at Vanderbilt University. From 2022 Summer to Present. Research Topic: Few-Shot Malware Classification. First Placement: Master of Science in Computer Science (MSCS) at Stanford University.

Teaching and Instructional Experience

- Graduate Teaching Assistant for CS3892: Projects in Computing for Sustainability at Vanderbilt University, *Undergraduate Capstone Project*, 2024 Spring. Instructor: Prof. Douglas H. Fisher.
- Graduate Teaching Assistant for CS3276/CS5276: Compiler Design at Vanderbilt University, 2023 Fall. Instructor: Prof. Kevin J. Leach.
- Community Teaching Assistant for Data Structure and Algorithm Training Camp (Part I and Part II) at Tsinghua University (MOOC), 2019 Spring. Instructor: Prof. Junhui Deng. [certificate]

- Online Course Mentor for Data Analysis Nanodegree at Udacity China, 2017 Spring. Instructors: Dr. Josh Magee, Ria Cheruvu & Matt Maybeno. **Outstanding Mentor Award** (Top 5% world-wide) [certificate]

Funds and Awards

Fellowships

- 2022 Research Fellowship from Defense Advanced Research Projects Agency (DARPA) (32500$/year)
- 2021 Research Fellowship from National Institutes of Health (NIH) ($36000/year)
- 2017 Roberto Roca Education Fellowship (Top 10 nationwide)
- 2015 Chinese National Fellowship for Overseas Studies (Full tuition fees & CA$6000/4 months)
- 2015 & 2016 & 2017 & 2018 First-class Scholarship at China University of Petroleum (CUP)
- 2014 Schlumberger Engineering Fellowship (Top 8 university-wide)
- 2013 Chinese National Scholarship for Outstanding Merits (Top 0.2% nationwide)

Honors and Awards

- 2020 First Runner-up in the 3rd JD.com Discovery Cup Patent Competition (Top 0.1% company-wide)
- 2020 Silver Medal Award of Distinguished Technical Recruiter at JD.com (Top 5% company-wide)
- 2020 Bronze Medal Award of Certified Technical Instructor at JD.com
- 2019 Beijing Outstanding Graduate Award (Top 0.1% nationwide)
- 2016 *Summa Cum Laude* for the Best Undergraduate Students at CUP (Top 1% university-wide)
- 2015 & 2017 Meritorious Winner of American Mathematical Contest in Modeling (Top 5% worldwide)
- 2013 & 2015 Third Prizes of Chinese National Petroleum Engineering Design Competition (Top 5% nationwide)