YUFAN ZHUANG

y5zhuang@ucsd.edu ◇ https://evanzhuang.github.io ◇ linkedin.com/in/yufan-zhuang

PROFILE

CS Ph.D. candidate (UC San Diego, expected 2025 Q4) with research experience at Apple Siri, AMD GenAI, Meta, Microsoft Research, and IBM Research (full-time pre-PhD).

Core expertise in LLM reasoning (continuous representations, agentic learning, long context understanding). 10+ peer-reviewed papers at top venues (ICLR, ACL, EMNLP, TMLR, FSE, ...), multiple patents, over 10K downloads on Huggingface, and open source software with 100+ stars.

EDUCATION

University of California San Diego, La Jolla, CA Sep 2021 - Present
PhD in Computer Science, Department of Computer Science & Engineering, Advisor: Prof. Jingbo Shang
Research Interests: Natural Language Processing, Large Language Models, Meta Learning

Columbia University, New York, NY Aug 2018 - Dec 2019
MS in Data Science, Data Science Institute, GPA: 3.96 / 4.00
Coursework: Machine Learning, Deep Learning, Reinforcement Learning, Mathematical Analysis

Hong Kong Polytechnic University, Kowloon, HK Sep 2014 - May 2018
BSc (Hons) with First Class Honors in Applied Mathematics, Minor in Computer Science, GPA: 4.00 / 4.00

EMPLOYMENT

Machine Learning Intern Jun 2025 - Sep 2025
Apple AIML - Siri *Cupertino, CA*

- Leading research on agentic systems for next-generation personal mobile assistants
- Developing realistic user simulation agents with automatic test case generation to improve Siri's robustness
- Designing evaluation metrics and benchmarking systems for agent performance in real-world mobile assistant scenarios
- Collaborating with cross-functional teams to integrate state-of-the-art LLM reasoning techniques into production systems

Research Scientist Intern Sep 2024 - Mar 2025
AMD GenAI *San Diego, CA*

- Pioneered agentic reasoning systems for long context understanding, resulting in state-of-the-art performance on HELMET.
- Developed ”Self-Taught Agentic Long Context Understanding” framework, enabling LLMs to autonomously improve their reasoning capabilities through iterative self-refinement (published at ACL 2025 Main Conference)

Research Scientist Intern June 2024 - Sep 2024
Meta - Reality Labs *Menlo Park, CA*

- Architected and pretrained Vision Language Models achieving competitive performance with more efficient SSM architecture
- Researched efficient VLM for high-definition images OCR, reducing inference latency by 50% while improving visual QA accuracy

PhD Research Intern June 2023 – Sep 2023
Microsoft Research, Deep Learning Group *Redmond, WA*

- Proposed “MetaTree”, a transformer-based decision tree algorithm outperforming XGBoost on 15+ tabular datasets
- Built pipeline to process 1M+ datasets for pretraining, creating the largest tabular meta-learning corpus
- Published in TMLR: Learning a Decision Tree Algorithm with Transformers

Graduate Research Assistant Sep 2022 - Current
UC San Diego, Department of Computer Science & Engineering, Advisor: Prof. Jingbo Shang *San Diego, CA*

- In pursuit of making LLMs more powerful

Research Software Engineer (Full-time) Jan 2020 - Sep 2021
IBM T. J. Watson Research Center, Supervisor: Dr. Alessandro Morari, Jim Laredo *New York, NY*

- Explored ways for neural methods to understand the logic structure of source code for better robustness and interpretability
- Researched novel graphical neural network architecture for vulnerability detection

- Published two papers & filed four patents (2 global patents, 2 US patents)

Graduate Research Intern May 2019 - August 2019
IBM T. J. Watson Research Center, Supervisor: Dr. Alessandro Morari, Jim Laredo *New York, NY*

- Designed and implemented framework for large scale data analysis on HPC
- Developed deep learning pipeline for vulnerability detection and localization

Graduate Research Assistant Jan 2019 - Dec 2019
Columbia University, Department of Computer Science, Advisor: Prof. Baishakhi Ray, Prof. Suman Jana *New York, NY*

- Worked on problems in software engineering that utilize NLP and deep learning, focusing on dataset bias analysis
- Developed models for automated vulnerability detection

Undergraduate Research Assistant Feb 2017 - Apr 2017 & Mar 2018 - Jul 2018
HKPU, Department of Applied Mathematics, Advisor: Prof. Xin Guo, Prof. Ting-kei Pong *Hong Kong*

- Built medical MRI demo using non-convex sparse optimization algorithm
- Conducted analysis of probabilistic as well as SVD-based topic modelling methods

PREPRINT & PUBLICATION

Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao “Text Generation Beyond Discrete Token Sampling” *arXiv preprint arXiv:2505.14827*, (2025).

Yufan Zhuang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Jingbo Shang, Zicheng Liu, Emad Barsoum “Self-Taught Agentic Long Context Understanding” *ACL’25*, (2025).

Yufan Zhuang, Chandan Singh, Liyuan Liu, Jingbo Shang, and Jianfeng Gao “Vector-ICL: In-context Learning with Continuous Vector Representations” *ICLR’25*, (2025).

Yufan Zhuang, Pierce Chuang, Yichao Lu, Abhay Harpale, Vikas Bhardwaj, and Jingbo Shang “Viper: Open Mamba-based Vision-Language Models” *https://huggingface.co/ViperVLM*, (2024).

Feng Yao*, **Yufan Zhuang***, Zihao Sun, Sunan Xu, Animesh Kumar, Jingbo Shang “Data Contamination Can Cross Language Barriers” *EMNLP’24*, (2024).

Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao “Learning a Decision Tree Algorithm with Transformers” *Transactions on Machine Learning Research (TMLR)*, (2024).

Yufan Zhuang, Zihan Wang, Fangbo Tao, Jingbo Shang “WavSpA: Wavelet Space Attention for Boosting Transformers’ Long Sequence Learning Ability” *NeurIPS UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Jim Laredo, Alessandro Morari, Udayan Khurana, “Incorporating Signal Awareness in Source Code Modeling: An Application to Vulnerability Detection” *ACM Transactions on Software Engineering and Methodology, Volume 32, Issue 6, Article No.: 145 pp 1–40*, 2023

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Jim Laredo, Alessandro Morari, Udayan Khurana, “Code Vulnerability Detection via Signal-Aware Learning” *IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 2023

Qiang Fu, **Yufan Zhuang**, Yushu Zhu, Xin Guo, “Sleeping Lion or Sick Man? Combining Computational Approaches to Deciphering Heterogeneous Images of Chinese in North America, 1978-2019.” *Annals of the American Association of Geographers (IF: 4.683)*, 2022

Yufan Zhuang, Sahil Suneja, Veronika Thost, Giacomo Domeniconi, Alessandro Morari, Jim Laredo “Software Vulnerability Detection via Deep Learning over Disaggregated Code Graph Representation.” *arXiv:2109.03341*, 2021

Sahil Suneja, Yunhui Zheng, **Yufan Zhuang**, Alessandro Morari, Jim Laredo “Towards Reliable AI for Source Code Understanding.” *ACM Symposium on Cloud Computing (SOCC) Vision Track*, 2021

Sahil Suneja*, Yunhui Zheng*, **Yufan Zhuang***(equal contribution)*, Alessandro Morari, Jim Laredo “Probing Model Signal-Awareness via Prediction-Preserving Input Minimization.” *ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2021

Qiang Fu, **Yufan Zhuang**, Jiaxin Gu, Yushu Zhu, Xin Guo, “Agreeing to Disagree: Choosing among Topic-Modeling Methods.” *Big Data Research (IF: 3.578)*, 2020

Luca Buratti, Saurabh Pujar, Mihaela Bornea, Scott McCarley, Yunhui Zheng, Gaetano Rossiello, Alessandro Morari, Jim Laredo, Veronika Thost, **Yufan Zhuang**, Giacomo Domeniconi, “Exploring Software Naturalness through Neural Language Models.” *arXiv:2006.12641*, 2020

Sahil Suneja, Yunhui Zheng, **Yufan Zhuang**, Jim Laredo, Alessandro Morari, “Learning to map source code to software vulnerability using code-as-a-graph.” *arXiv:2006.08614*, 2019

Qiang Fu, **Yufan Zhuang**, Jiaxin Gu, Yushu Zhu, Huihui Qin, Xin Guo, “Search for K: Assessing Five Topic-Modeling Approaches to 120,000 Canadian Articles.” *BPOD workshop at IEEE International Conference on Big Data: 3640-3647, 2019*

PATENT

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Alessandro Morari, Jim Alain Laredo, “Artificial intelligence model learning introspection”, US/WO Patent, No. US20230130781A1, 2023

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Alessandro Morari, Jim Alain Laredo, “Training data augmentation via program simplification”, US/TW/WO Patent, No. US20230113733A1, 2023

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Alessandro Morari, Jim Alain Laredo, “Complexity based artificial intelligence model training”, US/CN/JP Patent, No. US20230115723A1, 2023

Sahil Suneja, **Yufan Zhuang**, Yunhui Zheng, Alessandro Morari, Jim Alain Laredo, “Probing Model Signal Awareness”, US Patent, No. US20220358400A1, 2023

PROFESSIONAL ACTIVITIES

Reviewer: OOPSLA’20, TSE’21, AAAI’21, OOPSLA’21, NeurIPS’23 (UniReps), WWW’23, NeurIPS’24 (XAI), ICML’24, WWW’24, ICLR’25, TMLR, NeuRIPS’25

Teaching Assistant Experience:

ML/AI: CSE 250A (F’22, F’23, F’24), CSE 251A (S’23), CSE 151A (W’25), CSE 257 (W’23), CSE 150B (S’25)

Data Science: DSC 148 (W’24), DSC 258R (S’24)

Lab instructor of workshop, “An Introduction to Big Data and Automated Text Analysis for Social Scientists”, University of British Columbia, June 7-8, 2019

SELECTED ACCOMPLISHMENTS AND AWARDS

Jacobs School of Engineering Fellowship	2021
Department of Applied Mathematics Scholarship for Hall Residents	2017/18
The Hong Kong Polytechnic University (Eastern Canada) Association Scholarship	2017/18
The Hong Kong Polytechnic University Scholarship	2016/17
Honorable Mention, The Mathematical Contest in Modeling	2016
HKSAR Government Scholarship - Reaching Out Award	2015/16
Dean’s List	2014/15, 2016/17, 2017/18
Second Prize in National Olympiad in Informatics	2011