Yufan Zhuang

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

PROFILE

Research in LLM reasoning (continuous/agentic learning, long context understanding). First author of **Knowledge Flow**, which enabled open models to reach 100% on AIME-25 with no training/tools. Multiple first author papers at top venues (NeurIPS, ICLR, ACL, EMNLP, TMLR, FSE, ...), with US/Global patents.

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' Reasoning and Agentic Learning

Columbia University, New York, NY Aug 2018 - Dec 2019
MS in Data Science, Data Science Institute, GPA: 3.96 / 4.00

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

Research Intern, Microsoft CoreAI, Post-training Sep 2025 - Present

- Researching on infinite test-time scaling for LLM

Machine Learning Intern, Apple Siri 2025 Summer

- Lead research on agentic evaluation systems for next-generation personal mobile assistants
- Expected technical report on simulating realistic human users via finetuning GPT-4o

Research Scientist Intern, AMD GenAI 2024 - 2025

- Pioneered agentic reasoning systems for long context understanding, resulting in +14.7% on HELMET benchmark.
- First author paper AgenticLU published at ACL 2025 main conference (top NLP venue)

Research Scientist Intern, Meta Reality Labs 2024 Summer

- Pretrained efficient VLMs for high-definition OCR, reducing inference latency by 50% while improving general QA performance
- Architected and pretrained Viper vision language models, viper-mamba-7b and viper-jamba-52b

Research Intern, Microsoft Research, Deep Learning Group 2023 Summer

- Pretrained MetaTree, a transformer tabular model over 1M+ datasets for decision tree generation
- First author paper published at TMLR, 100+ stars on github, 10K+ total downloads on Huggingface

Research Engineer (Full-time), IBM T. J. Watson Research Center 2020 - 2021

- Explored ways for neural methods to understand the logic structure of source code for better robustness and interpretability
- Published 8 papers and 4 patents (2 global patents, 2 US patents)

Graduate Research Intern, IBM T. J. Watson Research Center 2019 Summer

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

PUBLICATIONS

LLM REASONING & AGENTIC LEARNING

Yufan Zhuang, Liyuan Liu, Dinghuai Zhang, Chandan Singh, Yelong Shen, Jingbo Shang, and Jianfeng Gao. "Knowledge Flow: Scaling Reasoning Beyond the Context Limit." *Notion Blog*, 2025.

Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao. "Text Generation Beyond Discrete Token Sampling." *NeurIPS*, 2025.

Yufan Zhuang, Chandan Singh, Liyuan Liu, Jingbo Shang, and Jianfeng Gao. "Vector-ICL: In-context Learning with Continuous Vector Representations." *ICLR*, 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*, 2025.

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

Feng Yao*, **Yufan Zhuang***, Zihao Sun, Sunan Xu, Animesh Kumar, Jingbo Shang. "Data Contamination Can Cross Language Barriers." *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. (*equal contribution)

EFFICIENT TRANSFORMER ATTENTION

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.

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.

AI FOR SOFTWARE ENGINEERING

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*, Vol. 32, No. 6, Article 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.

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***. "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. (*equal contribution)

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.

AI FOR SOCIOLOGY

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*, 2022.

Qiang Fu, **Yufan Zhuang**, Jiaxin Gu, Yushu Zhu, Xin Guo. "Agreeing to Disagree: Choosing among Topic-Modeling Methods." *Big Data Research*, 2020.

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*, pp. 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), CSE 101 (F'26)
Data Science: DSC 148 (W'24), DSC 258R (S'24)

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
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