

Zheng Huang

Phone: 434-328-0876 · Email: zheng.huang.gr@dartmouth.edu · [Homepage](#) · [Google Scholar](#)

Education

Dartmouth College	<i>Sept 2023 - Present</i>
PhD of Computer Science, GPA: 3.7/4, advisor: Prof. Yujun Yan	<i>Hanover, NH, United States</i>
University of Virginia	<i>Aug 2019 - Dec 2021</i>
Master of Computer Science, advisor: Prof. Jundong Li	<i>Charlottesville, VA, United States</i>
University of British Columbia	<i>June 2018 - August 2018</i>
UBC Visiting Student, Department of Computer Engineering	<i>Vancouver, BC, Canada</i>
Hebei University of Technology	<i>Sept 2015 - June 2019</i>
Bachelor of Computer Science, major GPA: 87/100	<i>Tianjin, China</i>

Research Interest

I am interested in **Large Language Models** and **Machine Learning (ML)**, including:

- **Large Language Models**: Exploring diffusion language models and instruction-tuned LLMs to investigate knowledge representation and learning dynamics, guided by neuroscience-inspired insights into intelligence and reasoning
- **Graph ML**: Designing frameworks for diverse graph data to enhance model generalizability and applying graph ML models to real-world applications, such as recommender systems

Research Experience

Amazon Web Services AI Lab	<i>June 2025 - Sept 2025</i>
<i>Applied Scientist Intern — Paper Under Review, Arxiv</i>	<i>Santa Clara, CA, United States</i>

- Investigated reasoning and uncertainty estimation in diffusion large language models (DLMs)
- Developed a novel MCTS framework based on token-level uncertainty to optimize DLMs initialization
- Designed a task-splitting strategy to decompose complex problems into subtasks and guide DLM generation
- Conducted in-depth studies on post-training, probabilistic sampling, and reasoning in language models

Small-World Organization in Large Language Models	<i>Sept 2025 - Present</i>
<i>Current Project, advised by Yujun Yan</i>	<i>Hanover, NH, United States</i>

- Investigating small-world organization in transformer attention networks, motivated by neuroscientific evidence that small-world topology underlies human intelligence, as a structural correlate of reasoning and intelligence
- Constructed inter- and intra-layer head-connectivity networks to examine local specialization and global integration across LLMs, revealing correlations with model fluid intelligence performance
- Using the small-world index as a quantitative signal to study post-training and reinforcement learning dynamics

Selected Publications

- **Zheng Huang**, Kiran Ramnath, Yueyan Chen, Aosong Feng, Sangmin Woo, Balasubramanian Srinivasan, Zhichao Xu, Kang Zhou, Shuai Wang, Haibo Ding, Lin Lee Cheong. “Diffusion Language Model Inference with Monte Carlo Tree Search”, Under Review, [Arxiv](#)

- **Zheng Huang**, Enpei Zhang, Yinghao Cai, Weikang Qiu, Carl Yang, Elynn Chen, Xiang Zhang, Rex Ying, Dawei Zhou, Yujun Yan. “Seeing Through the Brain: New Insights from Decoding Visual Stimuli with fMRI”, Under Review, [Arxiv](#)
- Weikang Qiu, **Zheng Huang**, Haoyu Hu, Aosong Feng, Yujun Yan, Rex Ying. “MindLLM: A Subject-Agnostic and Versatile Model for fMRI-to-Text Decoding”, ICML 2025, [Arxiv](#)
- **Zheng Huang**, Qihui Yang, Dawei Zhou, Yujun Yan. “Enhancing Size Generalization in Graph Neural Networks through Disentangled Representation Learning”, ICML 2024, [Arxiv](#)
- **Zheng Huang**, Jing Ma, Yushun Dong, Natasha Zhang Foutz and Jundong Li, “Empowering Next POI Recommendation with Multi-Relational Modeling”, SIGIR 2022. [Arxiv](#)
- Jing Ma, Yushun Dong, **Zheng Huang**, Daniel Mietchen and Jundong Li. ”Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US”, WWW 2022, [Arxiv](#)
- [Full Publication List \(Google Scholar\)](#)

Industry Experience

Alexa Speech Recognition, Amazon.com, Inc. *Mar 2022 - Jan 2023*

Machine Learning Engineer *Seattle, WA, United States*

- Developed Federated Learning (FL) systems to preserve users' privacy and improve the quality of speech recognition
- Worked on a team and delivered an on-device FL Recurrent Neural Network Transducer prototype that is capable of learning from the audio without relying on sending users voice recordings to the cloud
- Implemented FL on-device trainer that was constructed during the device idle time and deconstructed after the completion of training tasks to minimize the footprint of training in memory to avoid customer friction
- Conducted in-depth investigation on decentralized machine learning, privacy-preserving and AWS infrastructure

Technical Skills

- **Programming:** Python, Java, C++, C, R, Javascript, HTML, CSS, SQL, MATLAB
- **Tools:** PyTorch, TensorFlow, Scikit-learn, PySpark, AWS, Linux, Numpy, Pandas, Latex

Services

- **Conference reviewer:** ICLR 2025, NeurIPS 2025, ICLR 2024, NeurIPS 2024, ICLR 2023, NeurIPS 2024, ICLR 2023, JMLR 2022, ECML 2022, PAKDD 2021, WSDM 2021
- **Industry reviewer:** Amazon Machine Learning Conference 2022

Awards

- ICML Travel Grant 2024
- SIGIR Student Grant 2022
- Computer Science Department 2nd-level Academic Fellowship 2018
- Computer Science Department 1st-level Academic Fellowship 2017
- Computer Science Department 2nd-level Academic Fellowship 2016

Teaching

- Teaching Assistant, Deep Learning, Fall 2025
- Teaching Assistant, Deep Learning, Spring 2025