Zhuoyan Xu

Tel: (+1) 608-960-1191 | Email: zhuoyan.xu@wisc.edu | Homepage: http://zhuoyan-xu.github.io/

Education

University of Wisconsin-Madison — Madison, WI
Ph.D. in Statistics (Co-advised by: Yin Li, Yingyu Liang, Yiqiao Zhong) — Sep. 2020 - Dec. 2025 (expected)
M.S. in Computer Science — Sep. 2021 - May 2023

Wuhan University — China
B.S in Statistics — Aug. 2015 – Jun. 2019

Research Interests

My research focuses on the learning and adaptation of Foundation Model, including Large Language Models and Multimodal Models, with the goal of improving their capabilities and deployment efficiency.

Recent Publications

- **Zhuoyan Xu***, Khoi Duc Nguyen*, Preeti Mukherjee, Somali Chaterji, Saurabh Bagchi, Yingyu Liang, Yin Li. AdaLLaVA: Learning to Inference Adaptively for Multimodal Large Language Models. ICCV 2025. (* denotes equal contribution).
- **Zhuoyan Xu**, Debanjan Datta, Mukul Prasad. Augmenting Agent Memory With Temporal GraphRAG. Work done during internship at AWS summer 2025.
- Zidong Liu, **Zhuoyan Xu**, , Zhenmei Shi, Yingyu Liang. Can Language Models Compose Skills In-Context?. arXiv 2025.
- **Zhuoyan Xu**, Haoyang Fang, Boran Han, Bonan Min, Bernie Wang, Shuai Zhang. TabRAG: Efficient Table Retrieval and Understanding with Large Multimodal Models. Work done during internship at AWS summer 2024.
- Yingyu Liang*, Heshan Liu*, Zhenmei Shi*, Zhao Song*, **Zhuoyan Xu***, Junze Yin* Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers EMNLP 2025 Findings. (* denotes alphabetical order).
- Jiajun Song, **Zhuoyan Xu**, Yiqiao Zhong. Out-of-distribution generalization via composition: a lens through induction heads in Transformers. PNAS 2025.
- **Zhuoyan Xu**, Khoi Duc Nguyen, Preeti Mukherjee, Somali Chaterji, Yingyu Liang, Yin Li. AdaInf: Adaptive Inference for Resource-Constrained Foundation Models. ICML 2024 Workshop.
- **Zhuoyan Xu***, Zhenmei Shi*, Yingyu Liang. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability. COLM 2024. (* denotes equal contribution).
- Zhenmei Shi, Jenny Wei, **Zhuoyan Xu**, Yingyu Liang. Why Larger Language Models Do In-context Learning Differently? ICML 2024.
- **Zhuoyan Xu**, Zhenmei Shi, Jenny Wei, Fangzhou Mu, Yin Li, Yingyu Liang. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning. ICLR 2024.
- **Zhuoyan Xu**, Zhenmei Shi, Jenny Wei, Yin Li, Yingyu Liang. Improving Foundation Models for Few-Shot Learning via Multitask Finetuning. ICLR 2023 Workshop.

Professional Experience

Applied Scientist Intern — May 2025 – Aug. 2025
Amazon AWS AI — Bellevue, WA

- Augmenting Agent Memory With Temporal GraphRAG
- Develop **TempEval**, first comprehensive temporal reasoning evaluation for RAG system and agent memory.
- Propose **Astral**, a novel method that augments agent memory with temporal knowledge graphs, achieving superior temporal query performance while maintaining performance on standard queries

Applied Scientist Intern — May 2024 – Aug. 2024
Amazon AWS AI — Bellevue, WA

- Retrieval-augmented generation (RAG) for Multimodal Large Language Model
- Propose **TabRAG**, a novel framework that addresses table understanding challenges by directly utilizing table images in both retrieval and generation step
- Experimental validation is conducted using a newly constructed table image dataset from 14 public table understanding dataset, demonstrating the robustness and efficiency of our proposed framework

Machine Learning Engineer Intern May 2022 – Aug. 2022
John Deere Fargo, ND

- Developed a deep learning framework for electrical machine winding temperature prediction
- Architected an end-to-end data pipeline processing over 1M multivariate time series sequences, optimizing data quality and feature extraction
- Implemented ML models for time series prediction, delivering 50% accuracy improvement over baseline statistical approaches
- Created and deployed Cynet, a custom Python package that streamlined laboratory testing procedures

Research Assistant Sep. 2021 – present
UW - Madison Madison, WI

- Towards understanding and better adaptation of foundation models.

Academic Services

Conference Reviewer: NeurIPS 2024, ICML 2024-2025, ICLR 2025, AISTATS 2025, CVPR 2025, ICCV 2025

Preprints

- **Zhuoyan Xu**, Kris Sankaran. Spatial Transcriptomics Dimensionality Reduction using Wavelet Bases. F1000 Research.
- **Zhuoyan Xu**, Jiaxin Hu, Miaoyan Wang. Generalized tensor regression with covariates on multiple modes. arXiv.

Technical Skills

Languages: Python, R, Julia, Java, C, SQL, HTML
Developer Tools: Git, Linux, AWS, Azure, Docker, Google Cloud Platform, Slurm, LaTeX

Software

- **Xu, Z.**, Sankaran, K., 2022. R package *waveST*: Spatial Transcriptomics Dimensionality Reduction using Wavelet Bases. Published on GitHub. DOI: 10.5281/zenodo.6823315
- **Xu, Z.**, Hu, J. and Wang, M., 2019. R package *Tensorregress*: Generalized tensor regression with covariates on multiple modes. Published on the Comprehensive R Archive Network
- **Xu, Z.**, Solís-Lemus, C. 2022. Julia Package *HighDimMixedModels.jl*: Fitting mixed-effects models with high dimensional fixed effect variable.