ZHUOYAN XU

Tel: (+1) 608-960-1191 | Email: zhuoyanxu1@gmail.com | Homepage: http://zhuoyan-xu.github.io/

PROFESSIONAL EXPERIENCE

Amazon Web Services AI Lab Mar 2026 – Present
Applied Scientist II
- Build AI agents

EDUCATION

University of Wisconsin-Madison
Ph.D. in Statistics (Co-advised by: Yin Li, Yingyu Liang, Yiqiao Zhong) Sep 2020 - Jan 2026
M.S. in Computer Science Sep 2021 - May 2023

Wuhan University
B.S in Statistics Sep 2015 – Jun 2019

PH.D. THESIS

Towards Better Foundation Models: Theory and Methods for Adaptation and Deployment.

SELECTED PUBLICATIONS

- **Efficient Table Retrieval and Understanding with Multimodal Large Language Models**
  Zhuoyan Xu*, Haoyang Fang, Boran Han, Bonan Min, Bernie Wang, Shuai Zhang.* EACL 2026 Findings
- **AdaLLaVA: Learning to Inference Adaptively for Multimodal Large Language Models** ICCV 2025
  Zhuoyan Xu*, Khoi Duc Nguyen*, Preeti Mukherjee, Somali Chaterji, Saurabh Bagchi, Yingyu Liang, Yin Li*
  (denotes equal contribution).*
- **Can Language Models Compose Skills In-Context?** Preprint 2025
  Zidong Liu, ***Zhuoyan Xu****, Zhenmei Shi, Yingyu Liang*
- **Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers** EMNLP 2025 Findings
  Yingyu Liang, Heshan Liu*, Zhenmei Shi*, Zhao Song*,* ***Zhuoyan Xu****, Junze Yin**
  (denotes alphabetical order).*
- **Out-of-distribution generalization via composition: a lens through induction heads in Transformers**
  Jiajun Song, ***Zhuoyan Xu****, Yiqiao Zhong.* PNAS 2025
- **AdaInf: Adaptive Inference for Resource-Constrained Foundation Models** ICML 2024 Workshop
  Zhuoyan Xu*, Khoi Duc Nguyen, Preeti Mukherjee, Somali Chaterji, Yingyu Liang, Yin Li.*
- **Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability** COLM 2024
  Zhuoyan Xu*, Zhenmei Shi*, Yingyu Liang*
  (denotes equal contribution).*
- **Why Larger Language Models Do In-context Learning Differently?** ICML 2024
  Zhenmei Shi, Jenny Wei, ***Zhuoyan Xu****, Yingyu Liang*
- **Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning** ICLR 2024
  Zhuoyan Xu*, Zhenmei Shi, Jenny Wei, Fangzhou Mu, Yin Li, Yingyu Liang*
- **Improving Foundation Models for Few-Shot Learning via Multitask Finetuning**
  Zhuoyan Xu*, Zhenmei Shi, Jenny Wei, Yin Li, Yingyu Liang* ICLR 2023 Workshop
- **Foodie Traps within Facebook Cannabis Promotional Posts: Deploying Multimodal Deep Learning AIs to Monitor Audience Engagement**
  Linqi Lu, Xianshi Yu, ***Zhuoyan Xu****, Hyerin Kwon, Akhil P Reddy, Haohang Xin, Shifan Zhang, Ellie Fan Yang, Yin Li, Sijia Yang* Drug and Alcohol Dependence 2026

INTERNSHIP EXPERIENCE

Amazon AWS AI May 2025 – Aug 2025
Applied Scientist Intern
- Enhance 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

Amazon AWS AI May 2024 – Aug 2024
Applied Scientist Intern
- 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

John Deere May 2022 – Aug 2022
Machine Learning Engineer Intern
- Develop Deep Learning Models for Time Series Forcasting
- Architected an end-to-end data pipeline processing over 1M multivariate time series sequences, optimizing data quality and feature extraction
- Created and deployed Cynet, a custom Python package that streamlined laboratory testing procedures

China Merchents Bank May 2018 – Aug 2018
Data Scientist Intern
- Develop Machine Learning Models for Customer profiling

University of Wisconsin-Madison Sep 2021 – Jan 2026
Research Assistant
- Towards understanding and better adaptation of Foundation Models.

TEACHING EXPERIENCE

University of Wisconsin-Madison Sep 2020 – May 2022
Teaching Assistant
- STAT 371: Introductory Applied Statistics for the Life Sciences (Fall 2020)
- STAT 301: Introduction to Statistical Methods (Spring 2021, Summer 2021)
- STAT 303–305: R for Statistics (Fall 2021)
- STAT 479: Statistical Data Visualization (Spring 2022)

Instructor Feb 2020
- Introduction to Big Data

INVITED TALKS

Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning
IBM Research talk, Feb 2024
UW - Madison Student Seminar, Feb 2024

Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability
UW - Madison Student Seminar, Mar 2024

ACADEMIC SERVICES

Conference Reviewer: NeurIPS 2024, ICML 2024-2025, ICLR 2025, AISTATS 2025, CVPR 2025, ICCV 2025, ECCV 2026

TECHNICAL SKILLS

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