

ANZE XIE

(608)-236-3958 | a1xie@ucsd.edu | AnzeXie.github.io

EDUCATION

University of California-San Diego M.S. in Computer Science & Engineering	<i>Sept. 2022 - Jun. 2024</i> GPA: 3.9/4.0
University of Wisconsin-Madison B.S. in Computer Science & Statistics & Applied Mathematics	<i>Sept. 2018 - Dec. 2021</i> GPA: 3.99/4.00

WORKING EXPERIENCE

AI Engineering Intern	Institute of Foundation Models, USA
Large-scale foundation model training infrastructure <ul style="list-style-type: none">Conducted supervised finetuning based on Qwen2.5-72B using LLaMA-FactoryOptimized system parameters for efficient training of large-scale Mixture-of-Experts (MoE) models using Megatron-LMAssessed long-context training capabilities of Megatron-LM's context parallelism mechanism	<i>May. 2025 - Aug. 2025</i>
Pretraining data for coding capabilities <ul style="list-style-type: none">Collected, deduplicated, and cleaned a large-scale dataset of coding problems for pretrainingBuilt an efficient, scalable pipeline for generating synthetic coding problems using vLLM and RayGenerated and curated a dataset of more than 100 million high-quality synthetic coding problemsSynthesized multi-turn traces of coding agents to enhance pretraining data richness	<i>Aug. 2025 - Nov. 2025</i>

RESEARCH EXPERIENCE

Research Assistant <i>Advisor: Prof. Hao Zhang</i> Area of Research: Long-context LLM training systems and LLM evaluation	UCSD, USA <i>Mar. 2023 - May. 2025</i>
LongChat and LongEval <ul style="list-style-type: none">Curated a long-context dataset and proposed a finetuning method for extending LLM's context windowDeveloped an evaluation toolset to assess LLM's long-context capabilitiesFinetuned LLaMA-7B and LLaMA-13B models and extended the context window by 8xPublished on Instruction workshop at NeurIPS 2023	
DistFlashAttn <ul style="list-style-type: none">Developed a state-of-the-art memory-efficient attention mechanism optimized for long-context LLM trainingProposed a re-materialization-aware gradient checkpointing strategyConducted ablation study on the efficiency of contemporary systems, including Megatron-LM, DeepSpeed Ulysses, FlashAttention, and Ring attention.Enabled up to 8x longer sequences and achieved a 2x speedup in trainingPublished on COLM 2024	
GameArena <ul style="list-style-type: none">Developed an innovative platform that dynamically evaluates LLM reasoning capabilities through live computer gamesConducted retrospective analysis on LLM reasoning process to reveal specific LLM reasoning capabilitiesPublished on ICLR 2025	
Research Assistant <i>Advisor: Prof. Shivaram Venkataraman, Prof. Theodoros Rekatsinas</i> Area of Research: Systems for graph learning	UW-Madison, USA <i>Feb. 2021 - May. 2022</i>
Data Mining Over Paleobiology Database <ul style="list-style-type: none">Extracted a knowledge graph from a paleobiology relational database with SQL and PythonTrained graph embedding models over the knowledge graph and performed link predictionProvided insights on fact discovery for the team's paleobiologistPublished in VLDB 2021	

PUBLICATIONS

Lanxiang Hu*, Qiyu Li*, **Anze Xie***, Nan Jiang, Ion Stoica, Haojian Jin, and Hao Zhang. "[GameArena: Evaluating LLM Reasoning through Live Computer Games](#)." (ICLR 2025). Co-first authored.

Dacheng Li*, Rulin Shao*, **Anze Xie**, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. "[DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training](#)." (COLM 2024).

Dacheng Li*, Rulin Shao*, **Anze Xie**, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. "[How long can opensource llms truly promise on context length](#)." (Instruction workshop @ NeurIPS 2023).

Anze Xie, Anders Carlsson, Jason Mohoney, Roger Waleffe, Shanan Peters, Theodoros Rekatsinas, and Shivaram Venkataraman. "[Demo of marius: a system for large-scale graph embeddings](#)." *Proceedings of the VLDB Endowment* 14, no. 12 (2021): 2759-2762.

OPEN-SOURCE CONTRIBUTIONS

Contributed significantly to [LongChat](#) and [LongEval](#), a repository supports training and evaluation of long-context LLMs

Developed the data preprocessing, postprocessing, and rule-based configuration optimizer modules for [Marius](#) and [MariusGNN](#), a unified system for large-scale graph-learning tasks

TEACHING EXPERIENCE

Teaching Assistant
Scalable Data Systems
ML Systems

UCSD, USA
Jan. 2024 - Mar. 2024
Mar. 2024 - Jun. 2024

AWARDS

UW-Madison Undergraduate Scholarship for Summer Study, 2020, 2021

UW-Madison Dean's List (7 semesters)

SKILLS

Languages: Python, GO, C, C++, Java, SQL, Matlab, R, HTML/CSS, L^AT_EX

Frameworks and Tools: PyTorch, Tensorflow, Transformers, pytest, tox