

Zhuoxuan Zhang

332-250-5122 | zhuoxuan@cs.unc.edu | github.com/Zhuoxuan-Zhang

RESEARCH INTEREST

- Recommendation Systems: efficient, and robust models for long-sequence and multimodal recommendation.
- Agentic AI: multimodal reasoning frameworks for education, cognitive support, and health applications.
- Post-Training and RL Efficiency: efficient post-training methods and multimodal post-training scaling laws.
- Explainable AI: neuron-level and process-level interpretability for generative and recommendation models.

EDUCATION

University of North Carolina at Chapel Hill

Doctor of Philosophy in Computer Science

Chapel Hill, NC

Expected May 2029

Columbia University in the City of New York

Bachelor of Science in Computer Science

New York, NY

May 2024

- **Relevant Coursework:** Artificial Intelligence, Generative Models, Computer Architecture, Computer Networks, Operating Systems, UI Design, Algorithm Design, Database Systems, Cybersecurity

WORK EXPERIENCE

Research Scientist Intern, TikTok Feed Recommendation

May 2026 - Present

ByteDance Ltd.

- Developed long-sequence recommendation pipelines for large ranking models, supporting continuous training.
- Designed KV cache-optimization algorithms that achieved up to $192\times$ compression.

PUBLICATIONS / IN SUBMISSION (* INDICATES EQUAL CONTRIBUTION)

Diffusion-GR2: Diffusion Generative Reasoning Re-ranker

Zhuoxuan Zhang*, Kangqi Ni*, Yuhang Chen*, Mingfu Liang*, Xiaohan Wei, Yunchen Pu, Fei Tian, Chonglin Sun, Frank Shyu, Adam Song, Sandeep Pandey, Luke Simon, Tianlong Chen, Xi Liu
arXiv preprint.

Sparse Neurons Carry Strong Signals of Question Ambiguity in LLMs

Zhuoxuan Zhang, Jinhao Duan, Edward Kim, Kaidi Xu
EMNLP 2025.

The Koala Benchmarks for the Shell: Characterization and Implications

Evangelos Lamprou, Ethan Williams, Georgios Kaoukis, **Zhuoxuan Zhang**, Michael Greenberg, Konstantinos Kallas, Lukas Lazarek, Nikos Vasilakis
USENIX ATC 2025 (Best Paper Award).

GuideLLM: Exploring LLM-Guided Conversation with Applications in Autobiography Interviewing

Zhuoxuan Zhang*, Jinhao Duan*, Xinyu Zhao*, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, Kaidi Xu
NAACL 2025.

Demo: An Exploration of LLM-Guided Conversation in Reminiscence Therapy

Zhuoxuan Zhang*, Jinhao Duan*, Xinyu Zhao*, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, Kaidi Xu
GenAI4Health@NeurIPS 2024 (Best Demo Paper Award).

Scaling Laws for Multimodal Large Reasoning Models with Arbitrary Modality Combinations

Jie Peng, **Zhuoxuan Zhang**, Gengwei Zhang, Huaizhi Qu, Hossein Nourkhiz Mahjoub, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi Pari, Yanyong Zhang, Tianlong Chen
Under review.

Targeted Poisoning of Generative Recommendation Models via Context-Aligned Interactions

Zhuoxuan Zhang, Jinhao Duan, Xinyu Zhao, Jinduo Guo, Shuai Shao, Rui Li, Luke Simon, Zhun Deng, Tianlong Chen, Bo Sun

Under review.

Faithful Chains: Enforcing Temporal Consistency in CoT Reasoning via Bellman Regularization

Zhuoxuan Zhang, Zhen Tan, Dawei Li, Bohan Jiang, Song Wang, Huan Liu, Tianlong Chen

Under review.

RESEARCH EXPERIENCE

Diffusion Generative Reasoning Re-rankers

UNC, Meta

- Developed a block-diffusion reasoning re-ranker with conversion fine-tuning, on-policy distillation, and reinforcement learning to enable parallel decoding.
- Achieved $3.5\times$ higher decoding throughput while matching the autoregressive model's performance

Next-Generation Generative Recommendation Systems

UNC, Meta

- Developing scalable end-to-end generative recommendation models that unify multimodal understanding, long-term user modeling, and next-item generation.
- Studying robustness challenges in real-world deployment, including context-aware poisoning attacks and defense mechanisms specific to generative recommenders.

Multimodal Post-Training Scaling Laws for Large Reasoning Models

UNC, Honda Research Institute

- Studied how multimodal tokens, modality interactions, and post-training methods affect MLLM reasoning.
- Proposed a unified post-training scaling law for arbitrary modality combinations and reasoning types.
- Developed a modality-incremental framework revealing modality contributions and multimodal interference.
- Showed that token-balanced allocation improves post-training efficiency and stability.

Sparse Neurons Encode Ambiguity in LLMs

Brown University, Drexel University

Accepted at EMNLP 2025

- Led a neuron-level interpretability study revealing that question ambiguity is sparsely and linearly encoded in LLMs.
- Developed a probing and activation-steering pipeline to identify and control Ambiguity-Encoding Neurons, achieving over 90% accuracy and strong cross-domain generalization.
- Demonstrated that steering a small set of ambiguity-encoding neurons shifts model behavior from direct answering to abstention, outperforming all evaluated baselines.

LLM-Guided Conversation Frameworks

Columbia University, UNC, Drexel University, UT Austin

Accepted at NAACL 2025

- Designed GuideLLM, a conversational framework improving goal navigation, empathetic engagement, and long-context management in LLMs.
- Extended the framework to medical and health-chatbot applications, improving conversation relevance and user trust.

TEACHING EXPERIENCE

Teaching Assistant

Various Courses

Columbia University

- Taught systems-programming fundamentals, including shell scripting and C programming.
- Instructed more than 80 students in data structures and algorithms.
- Supported introductory computer science courses covering object-oriented Python and software-development best practices.

TECHNICAL SKILLS

Languages: Python, C/C++, Java, Swift

Deep Learning & ML: PyTorch, CUDA, Recommendation Systems, Generative Models, Adversarial Robustness, Neural Network Verification

Systems: Distributed Training, KV-Cache Optimization, Linux, Docker

Web & App Development: HTML/CSS, JavaScript, Django, Flask, Swift for iOS

Tools: Git, Heroku, LaTeX