Zhehao Zhang

Phone: (+1) 640-240-4027 ◇ Email: zhang.16420@osu.edu
Homepage: https://zhehaozhang123.github.io/
Google Scholar: https://scholar.google.com/citations?user=QG-BAGwAAAAJ&hl=en

EDUCATION

The Ohio State University — *Columbus, OH*
PhD. in Computer Science & Engineering, Supervisor: Prof. Yu Su & Prof. Huan Sun — *Sep 2025-Present*

Dartmouth College — *Hanover, NH*
M.S. in Computer Science — *Sep 2023-Mar 2025*

Shanghai Jiao Tong University (SJTU) — *Shanghai, China*
B.E. in Artificial Intelligence (Honor Class) — *Sep 2019-Jun 2023*

PUBLICATIONS

[1] **Zhang, Zhehao**, J. Chen, and D. Yang, "DARG: Dynamic evaluation of large language models via adaptive reasoning graph," *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] **Zhang, Zhehao**, W. Xu, F. Wu, and C. K. Reddy, "Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning," *Conference on Language Modeling (COLM)*, 2025.

[3] **Zhang, Zhehao**, R. A. Rossi, B. Kveton, *et al.*, "Personalization of large language models: A survey," *Transactions on Machine Learning Research (TMLR)*, 2025.

[4] C. Ziems, W. Held, O. Shaikh, J. Chen, **Zhehao Zhang**, and D. Yang, "Can large language models transform computational social science?" *Computational Linguistics*, 2024.

[5] **Zhang, Zhehao**, W. Ma, and S. Vosoughi, "Is gpt-4v (ision) all you need for automating academic data visualization? exploring vision-language models' capability in reproducing academic charts," *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[6] **Zhang, Zhehao**, R. Rossi, T. Yu, *et al.*, "Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use," *arXiv preprint arXiv:2410.16400*, 2024.

[7] **Zhehao Zhang**, Y. Gao, and J. Lou, "E^5: Zero-shot hierarchical table analysis using augmented LLMs via explain, extract, execute, exhibit and extrapolate," *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[8] **Zhang, Zhehao**, X. Li, Y. Gao, and J. Lou, "CRT-QA: A dataset of complex reasoning question answering over tabular data," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[9] **Zhang, Zhehao**, J. Chen, and D. Yang, "Mitigating biases in hate speech detection from a causal perspective," *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[10] **Zhang, Zhehao**, T. Yu, H. Zhao, K. Xie, L. Yao, and S. Li, "Exploring soft prompt initialization strategy for few-shot continual text classification," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] C. V. Nguyen, X. Shen, R. Aponte, Y. Xia, S. Basu, Z. Hu, J. Chen, M. Parmar, S. Kunapuli, J. Barrow, J. Wu, A. Singh, Y. Wang, J. Gu, F. Dernoncourt, N. K. Ahmed, N. Lipka, R. Zhang, X. Chen, T. Yu, S. Kim, H. Deilamsalehy, N. Park, M. Rimer, **Zhang, Zhehao**, H. Yang, R. A. Rossi, and T. H. Nguyen, "A survey of small language models," *arXiv preprint arXiv:2410.20011*, 2024.

[12] J. Wu, **Zhang, Zhehao**, Y. Xia, *et al.*, "Visual prompting in multimodal large language models: A survey," *arXiv preprint arXiv:2409.15310*, 2024.

[13] J. Wu, H. Lyu, Y. Xia, **Zhang, Zhehao**, J. Barrow, I. Kumar, M. Mirtaheri, H. Chen, R. A. Rossi, *et al.*, "Personalized multimodal large language models: A survey," *arXiv preprint arXiv:2412.02142*, 2024.

[14] Y. Xia, S. Mukherjee, Z. Xie, J. Wu, X. Li, R. Aponte, H. Lyu, J. Barrow, H. Chen, F. Dernoncourt, B. Kveton, T. Yu, R. Zhang, J. Gu, N. K. Ahmed, Y. Wang, X. Chen, H. Deilamsalehy, S. Kim, Z. Hu, Y. Zhao, N. Lipka, S. Yoon, T.-H. K. Huang, Z. Wang, P. Mathur, S. Pal, K. Mukherjee, **Zhang, Zhehao**, N. Park, T. H. Nguyen, J. Luo, R. A. Rossi, and J. McAuley, "From selection to generation: A survey of llm-based active learning," *arXiv preprint arXiv:2502.11767*, 2025.

[15] R. Luera, R. Rossi, F. Dernoncourt, A. Siu, S. Kim, T. Yu, R. Zhang, X. Chen, N. Lipka, **Zhang, Zhehao**, *et al.*, "Optimizing data delivery: Insights from user preferences on visuals, tables, and text," *arXiv preprint arXiv:2411.07451*, 2024.

RESEARCH EXPERIENCE

Research Intern Jun 2022 - May 2024
Stanford University, Supervisor: Prof. Diyi Yang Stanford, CA

- **Dynamic Evaluation of Large Language Models (LLMs) [1]:**
  - Built the **DARG** framework, introducing adaptive reasoning graphs to dynamically generate test data with controlled complexity, enhancing evaluation robustness for 15 SOTA LLMs across diverse reasoning tasks.
  - Benchmarked LLMs on DARG-generated data, revealing performance drops under increased complexity and demonstrating the framework's utility in improving LLMs through **fine-tuning with dynamic datasets**.
- **Bias Mitigation in Hate Speech Detection [9]:**
  - Conducted **causal analysis** to identify confounding factors in hate speech detection, introducing the **Relative Spuriousness** metric for evaluating spurious features and guiding effective bias mitigation.
  - Proposed **Multi-Task Intervention** and **Data-Specific Intervention** to mitigate spurious correlations in hate speech detection, achieving robust improvements across 9 datasets and enhanced OOD generalization.
- **LLMs for Computational Social Science (CSS) [4]:** Developed a roadmap for integrating LLMs into CSS, implemented prompting practices, and built fine-tuned models (e.g., T5, RoBERTa) to benchmark 13 LLMs on 24 tasks, showcasing their potential in augmenting human annotation and generative tasks.

Research Intern Oct 2023 – Feb 2024
Dartmouth College, Supervisor: Prof. Soroush Vosoughi Hanover, NH

- **Vision-Language Models (VLMs) for Automatic Data Visualization [5]:** Developed **AcademiaChart**, a dataset of 2525 **high-resolution** academic charts from LaTeX source code on arXiv, showcasing diverse AI conference visualizations. Benchmarked six VLMs for **code generation** to replicate these charts, using fine-grained human evaluations and automated metrics to highlight how SOTA closed-source VLMs (e.g., GPT-4-V) can **significantly reduce researchers' effort** in creating accurate and reusable visualizations.

INDUSTRY EXPERIENCE

Applied Scientist Intern Nov 2024 - Present
Amazon.com Services LLC - PXT CS Team Seattle, WA

- **LLM Over-Refusal Mitigation and Safety Alignment [2]:** Developed **FalseReject**, a large-scale dataset of 11K challenging prompts spanning 45 safety-related categories to benchmark and mitigate excessive refusals in LLMs. Conducted evaluations across 20 SOTA LLMs, revealing their overgeneralization tendencies in safety alignment. Implemented supervised fine-tuning and reinforcement learning techniques to calibrate model refusals, reducing unnecessary rejections while preserving helpfulness.

Research Scientist/Engineer Intern Jun 2024 - Sep 2024
Adobe Inc. - Data Science Lab, Mentor: Dr. Ryan A. Rossi San Jose, CA

- **Multi-Agent Framework for Enhanced Visual Perception VipAct:** Developed VipAct, an agent framework that enhances VLMs through **multi-agent collaboration** and **vision expert models**, enabling more precise visual perception and System-2 reasoning. It outperformed baseline methods in **visual perception and reasoning tasks**, excelling in **visual prompt** comprehension [12] and **multi-image** inference.
- **A Survey on Personalized LLMs [3]:** Conducted a comprehensive survey on the personalization of LLMs, bridging the gap between **text generation** and **downstream applications**. Introduced novel taxonomies for personalization techniques, granularity, evaluation metrics, and datasets. Formalized foundational concepts and identified critical challenges and open research directions.

Research Scientist Intern Dec 2022 - Aug 2023
Microsoft Research Asia - Data, Knowledge, Intelligence Lab Beijing, China

- **Hierarchical Table Analysis with Code-Augmented LLM-based Agent [7]:** Built E^5, a tool-augmented framework for **hierarchical table QA** using GPT-4, achieving **SOTA performance** with a **44.38 Exact Match improvement** while eliminating the need for hand-crafted exemplars. Developed F^3, an adaptive algorithm built on E^5, reducing token usage by **93%** to enable **efficient large-scale table analysis** with **limited-context LLMs** while preserving accuracy.
- **Complex Reasoning QA over Tabular Data (CRT-QA) [8]:** Developed **CRT-QA**, the **first** large-scale table QA dataset requiring multi-step complex reasoning, introducing a **detailed reasoning taxonomy**. Proposed **ARC (Auto-exemplar-guided Reasoning with Code)**, a tool-augmented language agent framework leveraging Python (Pandas), achieving **SOTA results** without handcrafted exemplars.

ACHIEVEMENTS

Graduate Fellowship, awarded by Ohio State University	*2025*
COLM 2025 Travel Grant	*2025*
ICLR Notable Reviewer	*2025*
Merit Scholarship, awarded by Dartmouth College	*2023-2025*
Zhiyuan Honor Scholarship and Merit Scholarship, awarded by SJTU	*2019-2023*

SKILLS

Programming Languages	Python, C/C++, MATLAB, JavaScript
Machine Learning Tools	PyTorch, Huggingface, Numpy, Scikit-learn, Pandas
LLM-related Tools	verl, vLLM, LangChain, LlamaIndex, Gradio, Ollama

SERVICE

Reviewer	EMNLP 2023, 2024; NeurIPS 2023, 2024, 2025; NAACL 2024; ACL 2024, 2025; COLM 2024 CIKM 2024, 2025; ICLR 2025; COLING 2025; IJCAI 2025; IEEE TNNLS Journal
Volunteer	EMNLP 2023; NAACL 2024