

# Zhehao Zhang

Language Agents · Agent Safety · LLM Alignment

✉ zhang.16420@osu.edu | 🌐 zhehaozhang123.github.io | 📄 Google Scholar | 🗣️ zzh-SJTU | 🌐 zhehao-zhang

## EDUCATION

- **The Ohio State University** 2025 – Present  
Ph.D. in Computer Science and Engineering (Advisor: Yu Su) Columbus, OH
- **Dartmouth College** 2023 – 2024  
M.S. in Computer Science Hanover, NH
- **Shanghai Jiao Tong University** 2019 – 2023  
B.Eng. in Artificial Intelligence (Honor Class) Shanghai, China

## RESEARCH EXPERIENCE

- **Netflix** Jun 2026 – Aug 2026  
Machine Learning Intern (Host: William Harvey) Los Gatos, CA
  - Building a **terminal-based agentic benchmark** for video cutting, with **agent harnesses** using **self-distilled agent skills** and **multi-agent orchestration** to elicit reflective cognitive behavior and prevent long-horizon reasoning collapse.
- **Amazon** Nov 2024 – Jun 2025  
Applied Scientist Intern (Hosts: Weijie Xu, Chandan K. Reddy) Seattle, WA
  - Built **FalseReject** (COLM 2025): a 16K-prompt resource across 44 safety categories to benchmark over-refusal in 29 SOTA LLMs; used SFT and RL to cut unnecessary refusals while preserving helpfulness. Featured on the [Amazon Science Blog](#). 🧐
  - Led **Distractor Injection Attacks** (COLM 2026): characterized reasoning-distraction vulnerabilities in large reasoning models (up to 60% accuracy drop) and built an SFT+RL defense improving robustness by 50+ points. 🧐
- **Stanford University – SALT Lab** 2022 – 2024  
Research Intern (Host: Diyi Yang) Stanford, CA
  - Developed **DARG** (NeurIPS 2024): an adaptive-reasoning-graph framework that dynamically generates test data with controlled complexity, exposing accuracy drops across 15 SOTA LLMs. 🧐
- **Adobe Research** Jun 2024 – Aug 2024  
Research Intern (Host: Ryan A. Rossi) San Jose, CA
  - Designed **VipAct** (AAAI 2026): a multi-agent framework coupling VLMs with vision-expert models for fine-grained visual perception; with GPT-4o it consistently outperformed all prompting and agent baselines across every sub-task on the BLINK and MMVP benchmarks.
  - Led **Personalization of LLMs: A Survey** (TMLR 2025), one of the most impactful surveys in the domain (250+ citations), introducing unifying taxonomies for personalized LLMs.
- **Microsoft Research Lab – Asia** Dec 2022 – Aug 2023  
Research Intern (Host: Yan Gao) Beijing, China
  - Engineered **E5** (NAACL 2024): a tool-augmented agent for hierarchical-table QA (+44.38 EM over prior SOTA), with an adaptive variant for token-limited settings 🧐; created **CRT-QA** (EMNLP 2023), a complex table-reasoning dataset 🧐.

## PUBLICATIONS

\*EQUAL CONTRIBUTION; 2000+ CITATIONS, FULL LIST ON [GOOGLE SCHOLAR](#)

- [1] Jaylen Jones\*, **Zhehao Zhang\***, Yuting Ning, Eric Fosler-Lussier, Pierre-Luc St-Charles, Yoshua Bengio, Dawn Song, Yu Su, Huan Sun. [When Benign Inputs Lead to Severe Harms: Eliciting Unsafe Unintended Behaviors of Computer-Use Agents](#). ICML, 2026.
- [2] **Zhehao Zhang**, Ryan A. Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, Nedim Lipka. [VipAct: Visual-Perception Enhancement via Specialized VLM Agent Collaboration and Tool-use](#). AAAI, 2026.
- [3] **Zhehao Zhang**, Weijie Xu, Shixian Cui, Chandan K. Reddy. [Distractor Injection Attacks on Large Reasoning Models: Characterization and Defense](#). COLM, 2026.

- [4] **Zhehao Zhang**, Weijie Xu, Fanyou Wu, Chandan K. Reddy. [FalseReject: A Resource for Improving Contextual Safety and Mitigating Over-Refusals in LLMs via Structured Reasoning](#). *COLM*, 2025.
- [5] **Zhehao Zhang**, Jiaao Chen, Diyi Yang. [DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph](#). *NeurIPS*, 2024.
- [6] **Zhehao Zhang**, Weicheng Ma, Soroush Vosoughi. [Is GPT-4V\(ision\) All You Need for Automating Academic Data Visualization? Exploring Vision-Language Models' Capability in Reproducing Academic Charts](#). *Findings of EMNLP*, 2024.
- [7] **Zhehao Zhang**, Yan Gao, Jian-Guang Lou. [E5: Zero-shot Hierarchical Table Analysis using Augmented LLMs via Explain, Extract, Execute, Exhibit, and Extrapolate](#). *NAACL*, 2024.
- [8] **Zhehao Zhang**, Tong Yu, Handong Zhao, Kaige Xie, Lina Yao, Shuai Li. [Exploring Soft Prompt Initialization Strategy for Few-Shot Continual Text Classification](#). *ICASSP*, 2024.
- [9] **Zhehao Zhang**, Xitao Li, Yan Gao, Jian-Guang Lou. [CRT-QA: A Dataset of Complex Reasoning Question Answering over Tabular Data](#). *EMNLP*, 2023.
- [10] **Zhehao Zhang**, Jiaao Chen, Diyi Yang. [Mitigating Biases in Hate Speech Detection from A Causal Perspective](#). *Findings of EMNLP*, 2023.
- [11] **Zhehao Zhang**, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, et al. [Personalization of Large Language Models: A Survey](#). *TMLR*, 2025.
- [12] Yuting Ning\*, **Zhehao Zhang**\*, Yash Kumar Lal, Boyu Gou, Junyi Li, Weitong Ruan, Chentao Ye, Rahul Gupta, Diyi Yang, Yu Su, Huan Sun. [SkillHarm: Lifecycle-Aware Skill-Based Attacks via Automated Construction](#). *arXiv preprint*, 2026.
- [13] Yuting Ning, Jaylen Jones, **Zhehao Zhang**, Chentao Ye, Weitong Ruan, Junyi Li, Rahul Gupta, Huan Sun. [When Actions Go Off-Task: Detecting and Correcting Misaligned Actions in Computer-Use Agents](#). *ICML*, 2026.
- [14] Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, . . . , **Zhehao Zhang**, et al. [From Selection to Generation: A Survey of LLM-based Active Learning](#). *ACL*, 2025.
- [15] Chien Van Nguyen, Xuan Shen, Ryan Aponte, . . . , **Zhehao Zhang**, et al. [A Survey on Small Language Models](#). *RANLP*, 2025.
- [16] Reuben Luera, Ryan A. Rossi, Franck Dernoncourt, . . . , **Zhehao Zhang**, Seon Gyeom Kim, Tak Yeon Lee. [Personalizing Data Delivery: Investigating User Characteristics and Enhancing LLM Predictions](#). *WWW Companion*, 2025.
- [17] Jian Xie, Tianhe Lin, Zilu Wang, Yuting Ning, Yuekun Yao, Tianci Xue, **Zhehao Zhang**, et al. [QUEST: Training Frontier Deep Research Agents with Fully Synthetic Tasks](#). *arXiv preprint*, 2026.
- [18] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, **Zhehao Zhang**, Diyi Yang. [Can Large Language Models Transform Computational Social Science?](#). *Computational Linguistics*, 2024.
- [19] Junda Wu, Hanjia Lyu, Yu Xia, **Zhehao Zhang**, et al. [Personalized Multimodal Large Language Models: A Survey](#). *arXiv preprint*, 2024.
- [20] Junda Wu, **Zhehao Zhang**, Yu Xia, Xintong Li, Zhaoyang Xia, et al. [Visual Prompting in Multimodal Large Language Models: A Survey](#). *arXiv preprint*, 2024.
- [21] Reuben Luera, Ryan Rossi, Franck Dernoncourt, . . . , **Zhehao Zhang**, Seon Gyeom Kim, Tak Yeon Lee. [Optimizing Data Delivery: Insights from User Preferences on Visuals, Tables, and Text](#). *arXiv preprint*, 2024.

## SKILLS

---

- **Programming:** Python, C/C++, JavaScript,  $\LaTeX$
- **Post-Training & Inference:** PyTorch, Jax, verl, vLLM, Harbor; SFT, RLHF/RLVR, OPD, LoRA
- **Agent Systems:** Agent harnesses (Claude Code / Codex), tool-use, multi-agent orchestration, MCP, Agent Skills
- **Research Areas:** LLM Safety & Alignment, Language Agents, AI Red-Teaming, LLM Evaluation, Synthetic Data

## HONORS & SERVICE

---

- **Graduate Fellowship, The Ohio State University** 2025
- **ICML Gold Reviewer; ICLR Notable Reviewer** 2025 – 2026
- **COLM 2025 Travel Grant** 2025
- **Zhiyuan Honor Scholarship & Merit Scholarship, SJTU** 2019 – 2023
- **Reviewer:** ICML, NeurIPS, ICLR, AAAI, ACL, EMNLP, NAACL, COLM, IJCAI, CIKM; TMLR, IEEE TNNLS