Zhuohao (Jerry) Zhang

zhuohao@uw.edu | +1 (206) 678-6079 | zhuohaozhang.com

I build systems that **collect, verify, and use human data** for AI, like agent-trace and expert-trajectory collection, human-labeled eval sets, and SFT data for multimodal models and computer-use agents. I also work on **scalable oversight** from the human side: how people check AI work they cannot verify directly.

Education

University of Washington | *Ph.D., Information Science* 2021 – Jun 2027 (expected)
University of Illinois Urbana-Champaign | *M.S., Computer Science*, GPA 4.0/4.0 2019 – 2021
Zhejiang University | *B.Eng., Computer Science with Honors*, GPA 3.87, top 5% 2015 – 2019

Related Experience

Anthropic | *AI Safety Research Fellow* Nov 2026 – Feb 2027

- Full-time AI safety research on frontier models with Anthropic researchers.

Apple | *Machine Learning Intern* Summers 2024 & 2025

- Built a dataset of **consequential agent actions**: designed an adversarial elicitation protocol and collection system where participants deliberately performed actions in mobile apps that would discomfort the phone owner if an agent did them (73% moderate/high impact vs. <8% in public datasets).
- Defined **8 action-impact evaluation dimensions** (e.g., reversibility, user intent, effects on others) and comprehensively evaluated frontier models on ∼1,500 double-annotated traces: the best model reached only 58% on overall impact, weakest on reversibility and user intent (paper).
- **Adopted in external agent-safety work:** Microsoft Research's Magentic-UI ActionGuard judge uses a prompt inspired by it; Apple's UINavBench uses it to exclude irreversible, high-impact tasks.

University of Washington | *Graduate Research Assistant, advised by Jacob O. Wobbrock* 2021 – Present

- **SlideAudit**: built a 2,400-slide eval set with known ground truth via **controlled flaw injection**, plus a crowd-labeling pipeline (145 annotators, 3 labels/slide) over a 27-type flaw taxonomy derived with 17 designers. Evaluated frontier LLMs as **design judges** across 24 configurations with a human-validated LLM grader; best macro-F1 **0.655** vs. 0.577 for prior work (paper).
- **DesignBench** (in progress): pipeline that turns expert design trajectories into post-training data for multimodal design agents. Designed **expert-trajectory collection** on 1,500 slide–task pairs with leakage-controlled splits; trained a **LoRA SFT** adapter for sketch-to-wireframe prediction; designed matched preference evals (expert-guided vs. task-only).

Microsoft Research | *Applied Scientist Intern* Summer 2023
Meta Reality Labs | *Research Scientist Intern* Summer 2022
Adobe Research | *Research Scientist Intern* Summer 2020

Honors

Anthropic AI Safety Research Fellowship, selected from 10,000+ applicants (<1%) 2026
Apple Scholars in AI/ML PhD Fellowship 2024
Student Lead, $4.6M federal grant on AI for accessibility 2024 – 2029
Best Paper Honorable Mention (CHI) and **Best Poster Honorable Mention** (UIST) 2022
13 peer-reviewed full papers, 8 first-authored (full list) 2019 – 2025

Skills

Post-training: SFT, LoRA, leakage-controlled splits, data quality control, human-preference evaluation.
Human data & evaluation: trace and trajectory collection, annotation pipelines, LLM-as-judge.
Engineering: Python, PyTorch, Hugging Face Transformers, TRL, PEFT, Vertex AI, JavaScript/TypeScript.