

Zhuohao (Jerry) Zhang

www.zhuohaozhang.com

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

Professional Summary

I build and evaluate **human-centered AI systems** that expand people's capabilities and preserve control over AI actions, focusing on agent safety, multimodal AI, and accessible authoring.

Education

University of Washington

Ph.D. in Information Science

2021 - now

Selected honors: Anthropic AI Safety Research Fellow (2026); Apple Scholars in AI/ML PhD Fellowship (2024); UIST Best Poster Honorable Mention (2022).

University of Illinois at Urbana-Champaign

M.S. in Computer Science

2019 - 2021

Selected honor: CHI Best Paper Honorable Mention (2022).

Zhejiang University

B.Eng. in Computer Science with Honors

2015 - 2019

Selected honor: ACM CHI Student Research Competition Winner (2019).

Relevant Work Experience

Anthropic

Incoming AI Safety Research Fellow

Starting Nov. 2026

- Will conduct research on the safety and alignment of frontier AI systems with Anthropic's alignment team.

Apple

Machine Learning Intern

Summers 2024 & 2025

- Developed a taxonomy of **UI-action impacts** with 12 domain experts and evaluated LLMs on 1,569 UI action examples, measuring their ability to anticipate consequences such as irreversible changes and effects on other users.
- Influenced safeguards and benchmarks for computer-use agents: Microsoft Research's [Magentic-UI](#) uses a system prompt inspired by this work in its ActionGuard judge, and Apple's [UINavBench](#) applies the taxonomy to exclude tasks with impact or reversibility concerns.

University of Washington

Graduate Research Assistant

2021 - now

- Created **SlideAudit**, a 2,400-slide benchmark for multimodal design evaluation, with human annotations and evaluation workflows spanning 24 model, prompt, and input configurations. Taxonomy-guided prompting raised GPT-4o macro-F1 from **0.476 to 0.655** in evaluation.
- Developed A11yBoard for blind users to explore and edit visual layouts, then A11yDeck for AI-assisted visual authoring and verification. A11yDeck lets creators inspect AI plans before execution and verify changes through measurable evidence and touch exploration. Serve as **student project lead** within UW CREATE's **\$4.6M RERC on AI for accessibility**, whose funded proposal included A11yBoard.
- Developed **DesignBench**, a benchmark and post-training pipeline built around 1,500 slide-task pairs to align multimodal design agents with expert intent. Implemented expert-trajectory collection, leakage-controlled dataset splits, and a LoRA supervised fine-tuning adapter for sketch-to-wireframe prediction. Designed matched human-preference evaluations of task-only and expert-guided outputs.

Relevant Work Experience	Microsoft Research <i>Applied Scientist Intern, EPIC & Ability Team</i> Summer 2023 • Developed ChartA11y, a smartphone app enabling blind users to explore charts through semantic navigation and direct touch mapping, with coordinated audio and tactile feedback. • Used 13 co-design sessions with two blind partners to refine chart navigation, trend exploration, and scatter-plot interactions.
Other Work Experience	Meta Reality Labs <i>Research Scientist Intern, Agios Team</i> Summer 2022 Researched how augmented reality and AI can help people with disabilities access contextual information about their surroundings. Adobe Research <i>Research Scientist Intern, VaaS Team</i> Summer 2020 Studied storytelling workflows with 15 business analysts and developed CODAS, a no-code tool for interactive data reports. Produced two first-author EuroVA papers.
Selected Publications	Z. Zhang, R. Chen, M. Zhong, and J. O. Wobbrock. <i>SlideAudit: A Dataset and Taxonomy for Automated Evaluation of Presentation Slides</i>. UIST 2025. Z. Zhang, E. Schoop, J. Nichols, A. Mahajan, and A. Swearngin. <i>From Interaction to Impact: Towards Safer AI Agent Through Understanding and Evaluating Mobile UI Operation Impacts</i>. IUI 2025. J. N. Li, Z. Zhang, and J. Ma. <i>OmniQuery: Contextually Augmenting Captured Multimodal Memories to Enable Personal Question Answering</i>. CHI 2025. Full publication list
Relevant Skills	AI evaluation and safety: End-to-end benchmark and evaluation design for multimodal models and UI agents: task construction, taxonomies and rubrics, annotation pipelines, model and prompt comparisons, error analysis, agent-action consequences, and approval policies. Post-training and human feedback: Expert demonstrations and trajectories; multimodal supervision, leakage-controlled splits, SFT/LoRA with TRL/PEFT, dataset quality control, and human-preference evaluation. Research engineering: Python, PyTorch, Hugging Face Transformers, TRL/PEFT, Vertex AI, and JavaScript/TypeScript; reproducible data, training, inference, and evaluation pipelines. Human-AI interaction: Human-in-the-loop control and verification; eliciting expert intent through participatory design, interviews, controlled user studies, and mixed-method evaluation.