Jiaxun Zhang

LinkedIn | Email: jiaxunz2@illinois.edu | (447) 902-1616

EDUCATION

University of Illinois at Urbana-Champaign Expected Graduation: May 2026
Bachelor of Science in Statistics & Computer Science GPA: 3.94/4.00
Relevant Coursework:
CS: Machine Learning, Natural Language Processing, Deep Learning for Computer Vision, Data Mining, Machine Learning Systems, Algorithms, Numeric Methods
Math & Statistics: Statistical Modeling, Applied Bayesian Analysis, Statistics and Probability, Linear Algebra

AWARD

- Dean's List – Fall 2022, Spring 2023, Spring 2024, Spring 2025
- Fiddler Innovation Undergraduate Student Fellowship Award, UIUC, 2025

RESEARCH INTEREST

• LLM/VLM Reasoning • Multi-agent Systems • Reinforcement Learning • LLM Agent Safety and Robustness
• Al for Science • AI for Healthcare and Mental Health

PUBLICATIONS

"*" indicates equal contribution

1. X Cao*, **Jiaxun Zhang***, B Li*, Q He*, J Chen, Z Chen, H Xu, J Cao, JM Rehg. "Reinforcing Agentic Medical Vision-Language Reasoner." Under Review at ARR
2. X Li*, A Juvekar*, **Jiaxun Zhang**, X Liu, M Wahed, K Nguyen, Y Shen, T Yu, I Lourentzou. "Counterfactual Segmentation Reasoning: Diagnosing and Mitigating Pixel-Grounding Hallucination." Under Review at CVPR 2026
3. K Zhu*, Z Liu*, B Li*, M Tian*, Y Yang*, **Jiaxun Zhang**, P Han, Q Xie, F Cui, W Zhang, X Ma, X Yu, G Ramesh, Y Su, J Wu, Z Liu, P Lu, J Zou, J You. "Where LLM Agents Fail and How They Can Learn from Failures." Under Review at ICLR 2026
4. Y Yuan, **Jiaxun Zhang**, T Aledavood, R Zhang, K Saha. "Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and a Relational Lens." Under Review at ACM CHI 2026
5. H Yu*, K Xuan*, F Li*, K Zhu, Z Lei, **Jiaxun Zhang**, Z Qi, K Richardson, J You. "TinyScientist: An Interactive, Extensible, and Controllable Framework for Building Research Agents." EMNLP 2025 (System Demonstrations)
6. **Jiaxun Zhang***, K Zhu*, Z Qi*, N Shang, Z Liu, P Han, Y Su, H Yu, J You. "SafeScientist: Toward Risk-Aware Scientific Discoveries by LLM." EMNLP 2025 (Main Conference, Recommend Oral)
7. H Jiang, X Hao, Y Huang, C Ma, **Jiaxun Zhang**, R Zhang. "Advancing Medical Radiograph Representation Learning: A Hybrid Pre-training Paradigm with Multilevel Semantic Granularity." ECCV 2024 Workshop

RESEARCH EXPERIENCE

Counterfactual Segmentation Reasoning: Diagnosing and Mitigating Pixel-Grounding Hallucination IL
NCSA SPIN Program Advisor: Prof. Ismini Lourentzou **June 2025 – Present**

- Formalized Counterfactual Segmentation Reasoning (CSR) and curated HalluSegBench with metrics evaluating grounding robustness and distinguishing vision- vs. language-driven hallucinations.
- Developed RobustSeg, reducing hallucinations by 30% and improving FP-RefCOCO(+/g) performance.

SciFy: Tool-Integrated Scientific Claim Verification Framework Urbana, IL
NCSA SPIN Program Advisor: Prof. Ismini Lourentzou **June 2025 – Present**

- Developed an end-to-end agentic reasoning framework integrating claim decomposition, retrieval, multi-turn reasoning, and verdict generation, enabling interpretable and verifiable scientific reasoning.

- Integrated retrieval search and MCP-based verification tools for grounded, tool-assisted claim verification.

SafeScientist: Toward Risk-Aware Scientific Discoveries by LLM Agents Urbana, IL
Advisor: Prof. Jiaxuan You **Jan 2025 – June 2025**

- First framework to enable comprehensive risk-aware scientific discovery, integrating multi-layer safety modules across prompt monitoring, agent collaboration, tool use, and ethical review.
- Proposed SciSafetyBench, a benchmark of 240 high-risk scientific tasks and 120 tool-specific risk scenarios across six domains, enabling the first systematic evaluation of scientific AI safety.
- Developed a multi-agent safety architecture that enables automatic role scheduling, inter-agent message auditing, and real-time ethical intervention, and designed SafeChecker, a hybrid semantic structural defense combining LLM-based risk classification via LLaMA-Guard and rule-based pattern detection, achieving 78.7% adversarial rejection across seven attack categories
- Achieved 35% higher safety performance and 44% improvement in ethical compliance over prior AI scientist systems, while maintaining comparable research quality under adversarial conditions.

MedVL-R1: Reinforcing Agentic Medical Vision–Language Reasoner Urbana, IL
Advisor: Prof. James Matthew Rehg **Jan 2025 – Aug 2025**

- First framework to enable tool-augmented multimodal reasoning for medical-vision language models through reinforcement learning, reducing hallucination and enabling interaction with external environments and tools.
- Designed and integrated a modular tool library for medical vision–language reasoning, incorporating segmentation, detection, and literature retrieval tools.
- Developed Tool-enhanced P-GRPO, a contrastive RL algorithm that rewards effective tool use and spatial reasoning robustness, improving cross-modal interpretability.
- Built two large-scale biomedical datasets (Multimodal-PMC-100M and MedVL-R1-CoT) supporting the first systematic CoT reasoning evaluation, improving accuracy by +15.8% on SLAKE and +8.5% on VQA-RAD.

Replika and Mental Health: Assessing the Emotional and Ethical Impacts of AI Companionship Urbana, IL
Advisor: Prof. Koustuv Saha **Jan 2025 – Sep 2025**

- First longitudinal causal study on AI companionship and mental health, analyzing 47K Reddit posts from 10K+ users with a stratified propensity-matching and Difference-in-Differences design to uncover mixed affective and cognitive shifts linked to chatbot engagement.
- Developed a multi-method analysis combining large-scale NLP affect modeling (LIWC, SVM-based classifiers), causal inference, and in-depth user interviews, to reveal the dual impacts of AI companionship on emotional wellbeing and dependence.

WORKING & TEACHING EXPERIENCE

Tsinghua University Beijing, China
Algorithm Engineer Intern **May 2023 – August 2023 & May 2024 – August 2024**

- Deployed and tested traffic light, obstacle, and LiDAR recognition models (Yolov5/7) on Ubuntu.
- Built CARLA-NVS dataset (10+ scenes, multimodal data with RGB, depth, LiDAR); reconstructed dense point clouds using COLMAP/MVS.
- Applied GS-Net to generate dense Gaussian ellipsoids, enhancing 3DGS initialization and rendering quality.

University of Illinois, Computer Science Department Urbana, Illinois
Course Assistant, CS 124 **Jan 2023 – May 2023**

- Assisted 50+ students with coursework, Java debugging, and the semester-long project.

SKILLS

- **Languages:** Python, R, Java, C++, JavaScript, LaTeX
- **Packages:** PyTorch, Hugging Face Transformers, NumPy, Pandas, scikit-learn