

XUCHENG YU

xy63@illinois.edu | +86 18762883800 | Urbana-Champaign, IL

Trustworthy AI • LLM Safety and Reliability • Agentic and Multi-Agent Systems

RESEARCH INTERESTS

Trustworthy foundation models, with a focus on adversarial robustness, auditable reasoning and evaluation, and reliable agentic and multi-agent decision-making under uncertainty.

EDUCATION

University of Illinois Urbana-Champaign (UIUC) Urbana-Champaign, IL | Aug. 2024 - Dec. 2025

Master of Engineering in Electrical and Computer Engineering (ECE) | GPA: 4.0/4.0

Guangdong Technion - Israel Institute of Technology Shantou, China | Sep. 2020 - Jul. 2024

Dual degrees awarded by GTIT and the Technion - Israel Institute of Technology

B.Sc. in Mathematics with Computer Science | GPA: 90.8/100, 3.7/4.0 | Dean's List, 2022-2023, 2023-2024

PUBLICATIONS AND MANUSCRIPTS

- **Xucheng Yu**, Haibo Jin, Huimin Zeng, and Haohan Wang. *SCI-Defense: Defending Manipulation Attacks from Generative Engine Optimization*. arXiv preprint arXiv:2605.21948, 2026. Submitted to NeurIPS 2026.
- Han Wang, Yifan Sun, Brian Ko, Mann Talati, Jiawen Gong, Zimeng Li, Naicheng Yu, **Xucheng Yu**, Wei Shen, Vedant Jolly, and Huan Zhang. *MonitorBench: A Comprehensive Benchmark for Chain-of-Thought Monitorability in Large Language Models*. Accepted at COLM 2026.
- **Xucheng Yu**, Emily Knox, and Haohan Wang. *Understanding Content Moderation in Large Language Models through Restricted Books: From Refusal to Warning*. Accepted at AIES 2026.
- Haibo Jin, Suijin Wang, **Xucheng Yu**, Haojing Luo, and Haohan Wang. *HEART: Harness Engineering in LLM Tool Use via Agent-Native Reusable Tool Primitives*. Submitted to NeurIPS 2026.
- Ke Chen, **Xucheng Yu**, Yufei Zhou, and Haohan Wang. *Prompt Stability Matters: Evaluating and Optimizing Auto-Generated Prompt in General-Purpose Systems*. Accepted at CPAL 2026.
- **Xucheng Yu**, Ruojia Tao, and Haohan Wang. *Closed-Loop Self-Improving Scientific Paper Generation: An Iterative Framework Combining AI Generation, Automated Review, and Targeted Optimization*. Manuscript.
- **Xucheng Yu**, Haibo Jin, Xiaoqi Wang, Runze Kong, and Haohan Wang. *Gap Detection for GEO Positioning and LLM Output Diversity*. Submitted to ICLR 2027.
- **Xucheng Yu**, Jiangping Chen, and Haohan Wang. *Rhetorical Distortion Detection across Platforms*. Submitted to The Web Conference 2027 (WWW 2027).
- Xiaoqi Wang, **Xucheng Yu**, Runze Kong, Haibo Jin, Huimin Zeng, Yifan Xu, and Haohan Wang. *GEO Survey and Unified Evaluation Benchmark*. Submitted to IEEE Transactions on Knowledge and Data Engineering (TKDE).

Submitted manuscripts are currently under peer review. Preprints are available on arXiv where applicable.

RESEARCH EXPERIENCE

Gap Detection for GEO Positioning and LLM Output Diversity Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang, Assistant Professor of ECE at UIUC Jul. 2026 - Oct. 2026

- Formulated a unified Gap Detection framework connecting three forms of AI content homogenization: LLM output diversity, GEO market positioning, and synthetic training-data collapse, using Demand x Defensibility optimization in embedding space.
- Developed theoretical and empirical analyses including the Unmet-Demand Theorem and Nash product objective; reduced LLM response collision by 12.8% on curated topics and 10.8% across hundreds of INFINITY-CHAT model responses.

Rhetorical Distortion Detection across Platforms

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

Jul. 2026 - Oct. 2026

- Designed a multi-platform, multi-language classifier for account-level rhetorical distortion across 332 accounts and 13,219 posts from RSS/Newsletter, YouTube, Bluesky, Reddit, Weibo, and Twitter/X.
- Built a three-tier pipeline combining rule-based signals, negative-pattern filtering, and GPT-4o-mini verification for five distortion dimensions; validated on 200 human-annotated posts with Cohen's kappa = 0.76 and a 12-point precision gain over a rule-only baseline.

GEO Survey and Unified Evaluation Benchmark

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

May 2026 - Oct. 2026

- Systematically characterized adversarial manipulation of LLM ranking systems across 20 generative engine optimization (GEO) pipelines, establishing a benchmark foundation for evaluating trustworthiness under deployment-time attacks.
- Compared AutoGEO, MAGEO, AgenticGEO, and E-GEO to identify algorithmic assumptions and new optimization and defense directions.

SCI-Defense: Semantic Manipulation Defense for LLM Ranking

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

Jan. 2026 - May 2026

- Proposed a three-component defense combining perplexity detection, Semantic Integrity Scoring, and Inter-Candidate Detection; achieved 1.000 precision and 0.000 false-positive rate across 1,200 Amazon ProductBench and MS MARCO evaluations.
- Developed six black-box attacks and showed that semantic manipulation through relevance inflation remains a structural blind spot for existing perplexity filters, safety classifiers, and paraphrasing defenses.

HEART: Reliable Multi-Agent Tool Use

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

Jan. 2026 - May 2026

- Designed Tool Primitives and HEART's Planner-Router-Verifier workflow; built ToolFace, a repository of 25,519 functions supporting dynamic retrieval, nested calls, and feedback-driven recovery.
- Evaluated on five diverse benchmarks, demonstrating a 10% improvement over SFT baselines and 6% over frontier commercial models, while reducing token consumption by up to 85%, highlighting sample efficiency in multi-agent tool use.

Content Moderation through Restricted Books

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

Jan. 2026 - May 2026

- Executed a 40,800-pair empirical study across 400 books and 17 prompt designs, evaluating six frontier models (Claude Sonnet 4.5, GPT-4o, Gemini 2.5 Flash, DeepSeek V3, Qwen-Plus, Grok-4.1-Fast); uncovered a near-zero refusal rate (0.07%), revealing systematic vulnerabilities in cross-model safety alignment.
- Characterized the shift from refusal to warning-based moderation, including warning gaps of 8-15 percentage points and prompt-dependent gaps reaching 19 points.

MonitorBench: A Comprehensive Benchmark for Chain-of-Thought Monitorability in Large Language Models

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Huan Zhang, Assistant Professor of ECE at UIUC

Jan. 2026 - Apr. 2026

- Contributed to MonitorBench, a comprehensive benchmark evaluating chain-of-thought monitorability across 1,514 instances, 19 tasks, and seven categories, with standard, direct-concealment, and monitor-aware-evasion settings.
- Implemented the impossible-coding-task evaluation adapted from ImpossibleBench and engineered a parallel multi-model inference and verification pipeline for large-scale monitorability experiments.

Prompt Stability and Multi-Agent Systems for Data Science

Urbana-Champaign, IL

Research Assistant | Advisor: Dr. Haohan Wang

Feb. 2025 - Sep. 2025

- Established semantic stability as a key reliability criterion for auto-generated prompts, devising an optimization framework linking prompt consistency to system-level task success.
- Architected a full-stack multi-agent platform with editable and regenerable agent dialogues, reaching 95% task completion and reducing user intervention by 40%.

RESEARCH SYSTEMS & PLATFORMS

Generative Engine Optimization Platform

Research Engineer | TopCited.ai

Urbana-Champaign, IL

May 2026 - Present

- Developing GEO SaaS features for AI visibility auditing, content optimization, and SCI-Defense-based ranking-manipulation detection across ChatGPT, Gemini, and Perplexity.

Natural-Language Tool Orchestration Platform

Research Engineer | Heelo

Urbana-Champaign, IL

May 2026 - Present

- Engineering natural-language tool orchestration with dynamic routing, multi-turn clarification, browser-based integrations, and credential-isolated execution.

PROFESSIONAL EXPERIENCE

Donovan's Piano Room

Full-Stack Engineer

Urbana-Champaign, IL

Feb. 2026 - Jul. 2026

- Resolved authentication and session-management defects across the Next.js frontend, Node.js/TypeScript backend, and Redis infrastructure by redesigning token-validation logic and decoupling protected routes from short-lived access tokens.

Bosch Automotive Products (Suzhou) Co., Ltd.

Software Engineering Intern

Suzhou, China

Sep. 2023 - Oct. 2023

- Developed and deployed an IoT data-processing and machine-learning pipeline using Airflow, Pandas, TensorFlow, FastAPI, and C/Cython, reducing preprocessing time by 40%, improving model accuracy by 12%, lowering query latency by 30%, and accelerating inference by 10×.

TECHNICAL PROJECTS

Augmented Reality Interactive Modeling for Kohler

Urbana-Champaign, IL | May 2025 - Aug. 2025

- Developed a real-time AR inspection pipeline using Python, Open3D, and Unity, achieving 0.1–0.2 mm precision and supporting ICP-based pose tracking for meshes with over 10 million vertices.

RLHF for Qwen2.5-7B-Instruct

Urbana-Champaign, IL | Jan. 2025 - May 2025

- Implemented an end-to-end PPO/GAE/KL-penalty pipeline with LoRA; achieved 70.1% pairwise preference accuracy and a 64.7% win rate versus 33.0% for the baseline on Stack Exchange data.

Distributed Consensus and Fault Detection

Urbana-Champaign, IL | Jan. 2025 - Apr. 2025

- Implemented Paxos and Raft over gRPC/Protobuf with 99.9% leader-election success and sub-200 ms election latency under network partitions; halved fault-detection time in simulated failures.

TEACHING EXPERIENCE

Peer Tutor & Grader

Guangdong Technion - Israel Institute of Technology

- Tutored lower-year students in mathematics and computer science coursework, supporting concept clarification, problem-solving strategies, and adaptation to a rigorous Technion-aligned curriculum.
- Graded assignments for Calculus, Topology, and Probability, providing feedback on mathematical reasoning, proof structure, and solution clarity.

TECHNICAL SKILLS

Research Methods: LLM evaluation, benchmark design, adversarial testing, large-scale empirical analysis, RLHF, multi-agent system evaluation

Programming Languages: Python, C/C++, Go, Java, TypeScript/JavaScript

Frameworks & Technologies: PyTorch, vLLM, SGLang, LoRA, FastAPI, Docker, React, Node.js, gRPC/Protobuf