

Xinjie Shen

✉ xinjie@gatech.edu

🌐 xinjie-shen.com

🎓 Education

Georgia Institute of Technology, Atlanta, USA

Aug. 2025 – Jun. 2029 (Expected)

Advisor: Pan Li *Ph.D. student* in Machine Learning focused on trustworthy agentic intelligence

South China University of Technology, Guangzhou, China

Sep. 2021 – 2025

Research Interests: Trustworthy Agentic Intelligence, Long-Horizon Post-Training, AI Safety and Privacy, Verifiable Environments

My research focuses on training long-horizon agents and studying how they respond to hidden intent and changing external states.

👤 Experience

Qwen | Post-Training Research Intern, Long-Horizon Agents

May 2026 – Oct. 2026

- **Long-Horizon Agentic Post-Training.** Contributed to agentic RL post-training from 35B models to flagship trillion-parameter models, building RL data and training pipelines for 35B- and 397B-parameter systems. Scaled real-reward RL for kernel optimization. Qwen3.8-Max-Preview reached No. 1 on NVIDIA's FlashInfer-Bench. Contributed to RSI runs exceeding 60 hours and 10,000 tool calls for chip design and inference-stack optimization.
- **Verifiable Hidden Dynamics Play (VHD-Play)** (*Qwen Technical Report, First Author*)
Introduced mechanism-first construction for scalable agentic RL, deriving hidden environment dynamics and verifiable outcome rewards from the same solved model. Generated 3,300 interactive environments where agents learn hidden mechanisms from partial observations and adapt their strategies through interaction. The learned behaviors transferred to eight unseen mechanism families.
- **E-Commerce Bench** (*Qwen Technical Report, Co-First Author*)
Co-built an open-source, 365-day evaluation of 18 frontier agents under evolving markets, repeated negotiation, delayed outcomes, and persistent memory demands.
- **MetaExp: Learning to Self-Improve through Experience via Verifiable Cross-Task Environment Scaling** (*Preprint, Co-Author*)
Trained agents to distill and reuse experience across streaming tasks without parameter updates. Improved transfer to unseen domains by 8.5 points on SpreadsheetBench and 14.8 on AutomationBench.

Open-Source Research | Agent Horizon, Long-Horizon RL Infrastructure

Built a veRL-based long-horizon RL framework that preserves environment state across context windows and supports asynchronous rollouts with episode-linked training. Designed to scale toward 1,000+ turns and 100M trajectory tokens.

AI Safety: Distributed-Intent Red Teaming and Closure-Aware Defense

Oct. 2025 – May 2026

Worked with Prof. Pan Li, Prof. Pin-Yu Chen, and Prof. Bo Li. Participated in SPAR with Yiyao Sun and Prof. Dawn Song.

- **SEAD: A State-Based Perspective on Attack and Defense in Tool-Using Agents** (*Preprint, First Author*)
Unified agentic attack and defense as partially observed state control. Developed DART and proactive SAGE for inspecting external state before consequential actions. SAGE reduced executable attack success from 48% to 4% online with verifiable evaluation.
- **One Turn Too Late: Response-Aware Defense Against Hidden Malicious Intent in Multi-Turn Dialogue** (*ICML 2026 AIWILD Workshop, Co-First Author*)
Studied hidden malicious intent emerging across multi-turn interactions. Built MTID and trained a 4B turn-level monitor on 16,000 trajectories with process supervision and RL. TurnGate reduced the miss rate by 20.2 percentage points.
- **The Trojan Knowledge: Bypassing Commercial LLM Guardrails via Harmless Prompt Weaving and Adaptive Tree Search** (*ICML 2026, Co-First Author*)
Showed that harmful intent can be distributed across benign-looking queries, evading turn-level safety checks. Built CKA-Agent with response-guided adaptive tree search, achieving over 95% guardrail bypass across commercial LLMs.

Privacy: Contextual Reasoning and Embodied Agents

Jul. 2025 – May 2026

Worked with Prof. Pan Li at Georgia Tech to study whether language and vision-language agents recognize privacy cues and use them when choosing actions.

- **Measuring Physical-World Privacy Awareness of Large Language Models** (*ICLR 2026, First Author*)
Built EAPrivacy, a four-tier benchmark with more than 400 scenarios across more than 60 scenes. Explicit reasoning sometimes hurt privacy performance, and models often chose neutral but unhelpful actions.
- **How Far Are VLMs from Privacy Awareness in the Physical World?** (*NeurIPS 2026, Co-First Author*)
Introduced ImmersedPrivacy, an interactive audio-visual benchmark that separates perception, privacy awareness, and agentic alignment. Across 12 VLMs, no model exceeded 65% accuracy under social-context shifts, and the best achieved 51% exact success when task completion conflicted with privacy.

Human-AI Interaction | Context Acquisition and Overreliance

Co-authored a CHI 2026 study of user overreliance on misleading LLM answers (77 participants). Independently built a benchmark of underspecified questions and found that LLMs often answer without seeking missing context.

Amazon Web Services | Research Intern, Sparse Retrieval

Nov. 2024 – Feb. 2025

- **Exploring ℓ_0 Sparsification for Inference-free Sparse Retrievers** (*SIGIR 2025, First Author*)
Developed at Amazon with Zhichao Geng and Dr. Yang Yang. Reduced model size while maintaining retrieval accuracy. The released models average 1.8 million monthly downloads.

Microsoft Research Asia | Research Intern, Autonomous R&D Agents

Jan. 2024 – Aug. 2024

- **Towards Data-Centric Automatic R&D** (*ICLR Workshop, Co-First Author*). Built benchmarks and evaluation metrics with Dr. Jiang Bian. The resulting RD-Agent project reached over 10,000 GitHub stars.
- **Collaborative Evolving Strategy for Automatic Data-Centric Development** (*Preprint, Core Author*)
Designed a collaborative strategy with knowledge-graph memory for iterative R&D. Received MSRA's "Star of Tomorrow" award.