

# Peizhi Niu

🌐 [Personal Website](#)   🎓 [Google Scholar](#)   ✉ [peizhin2@illinois.edu](mailto:peizhin2@illinois.edu)   ☎ (+1) 447-902-6992

## Education

University of Illinois Urbana-Champaign (UIUC), USA

PhD in Electrical and Computer Engineering

Sep 2024 – Present

- GPA: 4.0/4.0   Advisors: **Olgica Milenkovic** (IEEE Fellow); **Huan Zhang** (co-advisor since Aug 2026)
- Research Direction: LLM Agents, Agentic RL, Agent Safety, Trustworthy AI
- Expected graduation: **May 2028**

Shanghai Jiao Tong University (SJTU), China

Bachelor of Electronic Science and Technology

Sep 2020 – Jun 2024

- GPA: 90.52/100 (Major: 90.93/100)   Ranking: **Top 1 of 65** in department
- Honors: Outstanding Graduates in the City of Shanghai, National Scholarship

## Research Interests

- **LLM Agents**: Agentic RL, Self-Evolving Memory, Tool Use
- **Trustworthy AI**: Agent Safety, Machine Unlearning for Foundation Models
- **Collaborators**: Professors Pan Li, Dawn Song, Siheng Chen, Kamalika Chaudhuri, and Tong Zhang; Dr. Pin-Yu Chen (IBM Research)

## Industry Experience

Google, Remote

Student Researcher (Accepted to 2026 Fall Student Researcher Program)

Nov 2026 – May 2027

- Project: **Self-Synthesized RL Data for Scaling Agentic RL and Recursive Self-Improvement**
- Will study how a single LLM can **create its own RL training tasks**, aimed at the problems it currently fails, so that RL data can keep scaling without relying on a stronger teacher model
- Self-made tasks only help if their answers are correct and if training on them actually improves the model. Instead of checking every task in depth, which is expensive, we plan to check a **small random sample** carefully and use it to **correct a fast, rough estimate** applied to all tasks, keeping the data reliable at low cost
- Collaborators: Joy (Ruoqiao) Wei, Xidong Wu, Zheng Zhang

Amazon, Seattle, USA

Applied Scientist Intern

May 2026 – Aug 2026

- Project: **Self-Evolving Continual Learning Agent System** for consequential real-world enterprise tasks
- Built an agent that **accumulates reusable experience across sessions**, distilling completed task trajectories into a structured memory bank of procedures, actions, and lessons, and retrieving them to improve on later tasks without retraining
- Targeted enterprise workflows such as **file, project, and code management**
- Collaborators: Ge Shi, Vaibhav Gorde, Sarthak Ghosh, Christos Faloutsos, Min Cheng, Kevin Chen

## Research Highlights

LLM Agents: Memory, Tool Use, and Execution

Amazon, SJTU & UIUC, 2025 – Present

- **WASP-Mem.** Proposed Wiki-based Agent-Synthesized Procedural Memory, which consolidates execution traces and feedback into persistent, interlinked wiki pages for training-free continual learning. On TheAgentCompany, raised the checkpoint pass rate by **4.6–6.1 pp** over the strongest baseline across three models with **57–61% less storage**. A controlled study showed that human-defined schemas hurt performance, while one round of human diagnosis beats four rounds of self-judgment by **7.0–10.9 pp** → [OpenReview \(ICLR 2027 Submission\)](#)
- **MCP-Persona.** Constructed the first benchmark for evaluating LLM agents on real-world personalized MCP tools, spanning **24 MCP servers (12 personalized)** and **173 human-verified tasks** across social media, collaboration platforms, email, and content management. Proposed Tool-Traversal, a code-as-simulation paradigm that faithfully replicates account-bound tool behaviors (**94% accuracy, 0.87 METEOR**), revealing that even SOTA agents (GPT-5, Claude-Sonnet-4.5) struggle below 40% on personalized tool use → [ICML'26](#)
- **MCP-Flow.** Built an automated pipeline collecting data from **1,166 MCP servers and 11,536 tools**, producing 68,733 instruction-function-call pairs and 6,439 trajectories. Developed a web-agent-driven framework for large-scale MCP tool integration and model training → [ACL'26](#)
- **AgentSPEX.** Co-developed a YAML-based agent specification and execution language with explicit control flow, modular submodules, and a sandboxed harness for LLM agents. Outperformed compared baselines (CoT, ReAct, mini-SWE-

agent) on all **7 benchmarks** including SciBench, ChemBench, AIME 2025 (**100%**), and SWE-Bench Verified (**77.1%**) → [arXiv \(ARR Submission\)](#)

## Agent and LLM Safety

UIUC, 2024 – Present

- **SafeClawArena.** Proposed a computer-systems perspective on Claw-like agents and built a benchmark of **406 adversarial tasks** derived top-down from five classical cybersecurity principles across four attack surfaces (SSI, PSE, CDF, IPI). Evaluated **15 (platform, model) configurations** in containerized replicas of production agent platforms with canary-based taint tracking, exposing attack success rates of **20–70%** and **100%** for malicious plugins on every unhardened platform → [arXiv \(WWW 2027 Submission; Agents in the Wild Workshop\)](#)
- **TurnGate.** Developed a response-aware turn-level defender against multi-turn jailbreaks that detects the earliest harm-enabling turn via turn-level SFT + multi-turn RL on Qwen3-4B. Constructed the MTID dataset (16,000 dialogs with closure-turn annotations), achieving F1 **0.699** vs. 0.568 for the best baseline, with the RL stage reducing the miss rate from 37.9% to **17.7%** over the SFT warm start → [arXiv \(ICLR 2027 Submission\)](#)
- **Trojan Knowledge.** Developed a correlated-knowledge jailbreak agent using adaptive tree search, which decomposes harmful queries into seemingly innocuous prompts to bypass guardrails, achieving **95–99% attack success rates** on commercial LLMs and outperforming the strongest baseline by 15–21 percentage points → [ICML'26](#)
- **Temporal Memory Contamination.** Showed that memory accumulated over earlier, unrelated tasks raises the safety violation rate of memory-equipped LLM agents on later tasks. Introduced a trigger-probe protocol with a NullMemory counterfactual baseline, evaluated across three deployment scenarios, multiple memory architectures, and Claw-like agents (OpenClaw), and built a retrieval-time monitor that flags risk before generation (**0.97–0.98 recall**) → [arXiv \(ICLR 2027 Submission\)](#)

## Machine Unlearning

UIUC, 2024 – Present

- **GUARD.** Proposed a framework that assigns non-uniform unlearning weights based on data attribution scores, giving smaller weights to forget samples similar to retain data (preserving utility) and larger weights to dissimilar samples (ensuring effective forgetting) → [arXiv \(TMLR Submission\)](#)
- **Knowledge-Correlation Evaluation.** Designed an evaluation framework showing that seemingly “forgotten” facts can still be inferred through correlated knowledge in knowledge graphs, revealing that current evaluations overestimate unlearning effectiveness, with scores dropping by **at least 20%** in most settings (over 30% in more than half) → [NeurIPS'25](#)
- **DUMIC.** Co-developed a deep unlearning framework that discovers supporting reasoning paths via adaptive tree search with latent entities decoded from intermediate representations (J-Lens), then severs them with a confidence-weighted graph minimum cut, improving unlearning effectiveness by up to **359%** over superficial unlearning while preserving utility → [arXiv \(ICLR 2027 Submission\)](#)

## World Models for Robotic Planning

UIUC, 2026

- **Planning Sufficiency.** Co-developed a mechanism–response–decision sufficiency hierarchy characterizing what a world model must preserve for planning, showing that query-guided candidate proposals with query-independent, action-conditioned prediction improve held-out-objective regret while keeping predictions reusable across objectives → [arXiv \(ICLR 2027 Submission\)](#)

## Other Work: Graph Diffusion and Robustness

UIUC & CMU (Prof. Changliu Liu), 2023 – 2025

- **DMol.** Developed a molecule generation platform using discrete graph diffusion with a dynamic scheduling strategy, achieving **10x fewer** diffusion steps,  $\sim 2x$  shorter training, and  $\sim 40x$  faster sampling than DiGress → [NeurIPS'25](#)
- **SNR Certified Training.** Extended Signal-to-Noise Ratio into the robustness domain, addressing over-regularization in certified training → [TMLR'24](#)
- **ModelVerification.jl.** Developed a unified Julia-based framework integrating mainstream verification algorithms → [CAV'25](#)

## Publications

Published (\* equal contribution)

- [1] [\[ICML'26\]](#) R. Wei\*, **P. Niu\***, X. Shen\*, T. Tu, Y. Li, R. Wu, E. Chien, P.-Y. Chen, O. Milenkovic, and P. Li, “The Trojan Knowledge: Bypassing Commercial LLM Guardrails via Harmless Prompt Weaving and Adaptive Tree Search,” [ICML 2026](#).
- [2] [\[ICML'26\]](#) W. Wang\*, **P. Niu\***, G. Zou\*, X. Yang\*, J. Wang\*, H. Shi, Y. Du, J. Chai, X. Pang, S. Tang, Y. Wang, and S. Chen, “MCP-Persona: Benchmarking LLM Agents on Real-World Personal Applications via Environment Simulation,” [ICML 2026](#).
- [3] [\[ACL'26\]](#) W. Wang, **P. Niu**, Z. Xu, Z. Chen, J. Du, Y. Du, X. Pang, K. Huang, Y. Wang, Q. Yan, and S. Chen, “MCP-Flow: Facilitating LLM Agents to Master Real-World, Diverse and Scaling MCP Tools,” [ACL 2026](#).
- [4] [\[NeurIPS'25\]](#) **P. Niu**, Y. Wang, V. Rana, C. Rupakheti, A. Pandey, and O. Milenkovic, “DMol: A Highly Efficient and Chemical Motif-Preserving Molecule Generation Platform,” [NeurIPS 2025](#).
- [5] [\[NeurIPS'25\]](#) R. Wei\*, **P. Niu\***, H. H.-H. Hsu, R. Wu, H. Yin, M. Ghassemi, Y. Li, V. K. Potluru, E. Chien,

K. Chaudhuri, O. Milenkovic, and P. Li, “Do LLMs Really Forget? Evaluating Unlearning with Knowledge Correlation and Confidence Awareness,” [NeurIPS 2025](#).

[6] **[CAV’25]** T. Wei, H. Hu\*, L. Marzari\*, K. S. Yun\*, **P. Niu\***, X. Luo, and C. Liu, “ModelVerification.jl: A Comprehensive Toolbox for Formally Verifying Deep Neural Networks,” [CAV 2025](#).

[7] **[TMLR’24]** T. Wei, Z. Wang, **P. Niu**, A. Abuduweili, W. Zhao, C. Hutchison, E. Sample, and C. Liu, “Improve Certified Training with Signal-to-Noise Ratio Loss to Decrease Neuron Variance and Increase Neuron Stability,” [TMLR 2024](#).

**Under Review**

[8] **P. Niu**, G. Shi, V. Gorde, S. Ghosh, C. Faloutsos, M. Cheng, and K. Chen, “WASP-Mem: Let Agent Organize Procedural Memory. Human Spend Effort on Diagnosis,” Submitted to ICLR 2027, [OpenReview](#).

[9] **P. Niu\***, W. Qu\*, S. Gu, T. Shi, Y. Li, A. Tawaha, H. Alzahrani, V. Siu, B. Li, C. Wang, J. Zhang, B. Alomair, M. Jin, M. Chen, C. Wang, C. Spanos, and D. Song, “Understanding and Evaluating Claw-like Agent Security Through a Computer-Systems Lens,” Submitted to WWW 2027; accepted at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond, [arXiv](#).

[10] **P. Niu\***, E. Ma\*, H. Zhou, D. Zhou, H. Zhang, S. R. Etesami, and O. Milenkovic, “GUARD: Guided Unlearning and Retention via Data Attribution for Large Language Models,” Submitted to TMLR, [arXiv](#).

[11] X. Shen\*, R. Wei\*, **P. Niu\***, H. Wang, R. Wu, E. Chien, B. Li, P.-Y. Chen, and P. Li, “One Turn Too Late: Learning When to Intervene Against Multi-Turn Malicious Intent,” Submitted to ICLR 2027, [arXiv](#).

[12] R. Wei, H. H.-H. Hsu, **P. Niu**, Y. Li, and P. Li, “What Must a World Model Distinguish for Planning?” Submitted to ICLR 2027, [arXiv](#).

[13] J. Wang, **P. Niu**, H. Yin, H. H.-H. Hsu, P. Li, and R. Wei, “Making LLMs Truly Forget: Deep Unlearning by Searching, Selecting, and Severing Knowledge Paths,” Submitted to ICLR 2027, [arXiv](#).

[14] A. Al-Tawaha, S. Gu, **P. Niu**, R. Jia, and M. Jin, “Remembering More, Risking More: Longitudinal Safety Risks in Memory-Equipped LLM Agents,” Submitted to ICLR 2027, [arXiv](#).

[15] P. Wang\*, J. Huang\*, J. Yao\*, R. Pan, **P. Niu**, Y. Liu, R. Wang, R. Lu, Y. Guo, and T. Zhang, “AgentSPEX: An Agent SPecification and EXecution Language,” Submitted to ACL Rolling Review (ARR), [arXiv](#).

**Awards and Honors**

|                                                     |      |
|-----------------------------------------------------|------|
| • NeurIPS 2025 Travel Award                         | 2025 |
| • Outstanding Graduates in the City of Shanghai     | 2024 |
| • National Scholarship of China (Top 1% nationwide) | 2022 |
| • First-class Scholarship of SJTU                   | 2022 |
| • Excellent Student Cadre of SJTU                   | 2022 |

**Services**

- **Program Committee or Reviewer:** NeurIPS, ICML, ICLR, TMLR
- **Leadership:** Student Union of School of Electronic Information and Electrical Engineering, Department Secretary

**Skills**

- **Programming Languages:** Python, C/C++, Julia, MATLAB, VHDL
- **ML Frameworks:** PyTorch, JAX, Hugging Face Transformers
- **Tools:** Git, Docker, LaTeX, VS Code, Linux