Shuai Shao

Shanghai | shaoshuai.ederson@sjtu.edu.cn | 15620352810 | Google Scholar

Education

Shanghai Jiao Tong University French & Information Engineering Sept 2022 – Present

- **GPA**: 3.8/4.3
- **Coursework:** Probability and Statistics: 99, Topology: 98, Mathematical Foundation of AI: 95, Program Design and Algorithm Analysis: 95, Advanced Algebra: 94, Integration Series and Fourier Analysis: 92.
- **Award:** SJTU-SPEIT First-Class Academic Scholarship

Research Interests: RL for LLMs and Agents, Data Synthesis, LLM-Agent Safety & Alignment, RL Theory

Experience

Research Intern, Shanghai AILab, Coadvised by Dr. Dongrui Liu and Dr. Jing Shao Mar 2025 – Jan 2026

- As a **co-first author**, led the construction of **RiOSWorld** (Accepted by **NeurIPS 2025**)—the first risk assessment benchmark for multimodal computer-use agents (**1,500+ monthly downloads** on Hugging Face). The benchmark comprises 492 risk tasks spanning six application domains. Proposed a dual-dimension taxonomy of "user-sourced risks" and "environment risks," and designed an evaluation framework covering risk intent recognition to risk goal completion.
- Contributed as a **key contributor** to the **SafeWork-R1** technical report (WAIC 2025). Responsibilities included: (1) Efficient reasoning **RL post-training** on models ranging from **7B to 72B**—jointly optimizing safety and reasoning efficiency of large reasoning models through reinforcement learning; (2) Safety and efficient reasoning evaluation on benchmarks (MM-SafetyBench, MSSBench, SIUO, etc.), validating the AI-45° Law for co-evolving safety and capability.
- As **first author**, **proposed the concept of "Misevolution"** and led the first systematic study on misevolution risks in self-evolving LLM agents (accepted by **ICLR 2026**). Investigated four evolution pathways—model fine-tuning, memory accumulation, tool creation, and workflow optimization—revealing emergent risks such as safety alignment degradation and tool vulnerabilities even in agents built on top-tier LLMs like Gemini-2.5-Pro.
- As one of **Student Lead** for the **AgentDoG** project (∼**300 GitHub stars**, **#1 Hugging Face Daily Paper** on release day), responsible for large-scale data synthesis and benchmark construction. Developed a taxonomy-guided agent risk trajectory synthesis pipeline and contributed to the ATBench, achieving SOTA performance on R-Judge and ASSE-Safety benchmarks for agentic safety diagnosis and guardrail.

Visiting Student, Shanghai Innovation Institute, Advised by Prof. Weinan Zhang Sept 2025 – Present

- **Lead the development** of **MonoScale**, an expansion-aware update framework for scaling LLM-based multi-agent systems (MAS). Addressed the router cold-start problem when integrating new agents by proposing agent-conditioned task customization and natural-language memory updates.
- Formalized sequential agent augmentation as a contextual bandit and proved a **monotonic non-decreasing performance guarantee** via trust-region memory optimization, providing theoretical safety bounds for open-ended MAS expansion.
- Demonstrated stable performance gains on GAIA and Humanity's Last Exam as the agent pool scales, outperforming naive scale-up and strong-router baselines (e.g., GPT-5, Gemini-2.5-Pro).

Research Intern, APEX Lab, SJTU, Advised by Prof. Weinan Zhang Oct 2024 – Sept 2025

- Contributed to the development of **AgentNet** (Accepted by **NeurIPS 2025**)—a decentralized multi-agent coordination framework for LLM-based agents. Key contributions: (1) Designed a retrieval-based memory system enabling real-time skill refinement and specialization, allowing agents to efficiently adapt to diverse task types (coding, math, QA, etc.); (2) Implemented the self-evolving routing mechanism, enabling agents to autonomously adjust connections and routing based on task requirements, eliminating the need for a central coordinator. Experiments show AgentNet outperforms single-agent and centralized multi-agent baselines in task accuracy.

Research Intern, IIOT Lab, SJTU, Advised by Prof. Jiaxin Ding Mar 2024 – Oct 2024

- As **co-second author**, contributed to the **Extreme Value Policy Optimization (EVO)** algorithm (Accepted by **ICML 2025**). Addressed the limitation of expectation-based constraints in safe RL that ignore extreme "black swan" events by designing a risk-aware constrained RL algorithm based on Extreme Value Theory (EVT). Key

contributions: (1) Proposed an extreme value quantile optimization objective to explicitly capture extreme samples in the cost tail distribution; (2) Designed an extreme-priority replay mechanism to amplify learning signals from rare but high-impact samples; (3) Theoretically proved upper bounds on constraint violations during policy updates. Experiments show significant reduction in constraint violation probability and variance compared to SOTA baselines (PPO-Lagrange, CPO, WCSAC, QCPO, etc.) while maintaining policy performance.

Publications

* Equal Contribution † Corresponding Author

Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents ICLR 2026

Shuai Shao[*], Qihan Ren[*], Chen Qian, Boyi Wei, Dadi Guo, Yang JingYi, Xinhao Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu†, Jing Shao†

MonoScale: Scaling Multi-Agent System with Monotonic Improvement Under Review

Shuai Shao, Yixiang Liu, Bingwei Lu, Weinan Zhang†

RiOSWorld: Benchmarking the Risk of Multimodal Computer-Use Agents NeurIPS 2025 Poster

Jingyi Yang*, **Shuai Shao***, Dongrui Liu†, Jing Shao†

SafeWork-R1: Coevolving Safety and Intelligence under the AI-45 Law Technical Report

Shanghai AILab

Extreme Value Policy Optimization for Safe Reinforcement Learning ICML 2025 Poster

Shiqing Gao, Yihang Zhou*, **Shuai Shao***, Haoyu Luo, Yiheng Bing, Jiaxin Ding†, Luoyi Fu, Xinbing Wang

AgentNet: Decentralized Evolutionary Coordination for LLM-based Multi-Agent Systems NeurIPS 2025 Poster

Yingxuan Yang*, Huacan Chai*, **Shuai Shao**, Yuanyi Song, Siyuan Qi, Renting Rui, Weinan Zhang†

Towards Self-Evolving Benchmarks: Synthesizing Agent Trajectories via Test-Time Exploration ICLR 2026

Dadi Guo*, Tianyi Zhou*, Dongrui Liu, Chen Qian, Qihan Ren, **Shuai Shao**, et al.

The Why Behind the Action: Unveiling Internal Drivers via Agentic Attribution Under Review

Chen Qian*, Peng Wang*, Dongrui Liu, Junyao Yang, Dadi Guo, **Shuai Shao**, et al.

AgentDoG: A Diagnostic Guardrail Framework for AI Agent Safety and Security Technical Report

Shanghai AILab

Skills & Additional Information

Service: Reviewer for ICML 2026, NeurIPS Responsible FM Workshop, AAAI LaMAS Workshop

Talks: Invited Talk at NICE Community

Media Coverage: “Your Agent May Misevolve” covered by ; “RiOSWorld” covered by

Languages: Chinese(Native), English(**GRE: V155/Q170/AW3.5**), French(**DELF B2**)

Extracurricular Activities: Vice Principal Violist in the High-Level Art Troupe and **Main Goalkeeper** for the Academy Soccer Team at Shanghai Jiao Tong University