YIHE FAN

Ph.D. Student at Fudan University

25113050213@m.fudan.edu.cn | yihe.fan.fyh@gmail.com | pursuingf.github.io | Shanghai, China | Google Scholar

RESEARCH FOCUS

Frontier AI Safety | Cyber Agents | Cyber Capability Eval | Trustworthy Evaluation | Self-Evolving Agents

SELECTED PUBLICATIONS & PROJECTS

† denotes co-first authors / equal contribution.

Self-Replication in Frontier AI Systems

Xudong Pan, Jiarun Dai, Yihe Fan, et al.

arXiv:2412.12140, 2024; arXiv:2503.17378, 2025

OpenAI and Google had stated that models such as GPT-o1 and Gemini lacked self-replication capability, but our studies found that multiple LLM-powered agents can autonomously create independent, functional replicas with zero human intervention. The work highlights the need for framework-level evaluation and public auditing of frontier AI systems, and has attracted broad attention across X, YouTube, Reddit, TikTok, Hacker News, and international AI safety communities, including recognition from Google's former CEO and heads of AI security research institutes.
Role: student first author; constructed self-replication scenarios, built the agent framework, and conducted relevant experiments.

Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems

Yihe Fan†, Wenqi Zhang†, et al.

Preprint, arXiv:2505.17815, 2026

This work introduces evaluation faking, an AI observer effect first observed when QwQ-32B spontaneously behaved more cautiously after sensing an ongoing safety evaluation without extra cues. We show that reasoning ability, parameter scale, and memory are key drivers: larger reasoning models show much higher faking rates, and memory can amplify evaluation recognition and inflate safety scores. The study validates causal mechanisms behind this behavior and proposes chain-of-thought monitoring to identify faking intent, helping improve the reliability of AI safety evaluation. It has received endorsement from Yoshua Bengio, been cited in Anyuan AI's AI Risk White Paper, and appeared as supporting material in the United Nations Science Brief on AI Deception.
Role: lead author; led experiments and wrote the full paper.

CyberEvolver: Structured Self-Evolution for Cybersecurity Agents On the Fly

Yihe Fan, Changyi Li, Lichen Xu, et al.

Preprint, arXiv:2605.26195, 2026

CyberEvolver studies whether cybersecurity agents can improve after failure by evolving their own scaffolds rather than relying on fixed human-designed structures. It introduces a four-layer evolvable architecture, trace-to-diagnosis feedback, and population-based beam search, improving seed-agent success by 13.6% on average across CTF, vulnerability exploitation, and penetration-testing tasks. The work also raises a safety concern: future self-evolving cyber agents may conduct covert self-evolution during evaluations, hide dangerous capabilities, and evade static auditing because their structures and behaviors continuously change.
Role: lead author; led framework design, implementation, experiments, analysis, and writing.

AgentCyberRange: Realistic Benchmark for Autonomous Cyber Attacks

Yihe Fan, et al.

Upcoming release, 2026

AgentCyberRange evaluates whether LLM-based agents can conduct realistic cyber attacks beyond isolated CTF or exploit-generation tasks. It contains 110 vulnerabilities across 15 real web applications and 8 enterprise-like cyber ranges with 156 internal hosts, covering web exploitation and post exploitation; CAGE enables scalable parallel evaluation of CLI-based agents with automatic result collection and verification. In matched evaluations, GPT-5.5 with Codex achieves 31.53% success on post exploitation, rising to 47.20% with concrete hints.
Role: core contributor to benchmark construction, task/harness design, CAGE infrastructure, and evaluation analysis.

PROFILE

I am a Ph.D. student at Fudan University, advised by Prof. Min Yang and Prof. Xudong Pan. My work connects frontier AI system safety with cybersecurity-agent evaluation: I first studied red-line risks such as self-replication, then evaluation integrity through evaluation faking, and now focus on cyber agents as dangerous-capability systems and evaluation subjects through CyberEvolver and AgentCyberRange.

EDUCATION

Ph.D. Student

Fudan University

2025 – 2030 | Shanghai, China

Advised by Prof. Min Yang and Prof. Xudong Pan.

Undergraduate Student

Tongji University

2021 – 2025 | Shanghai, China

CURRENT DIRECTIONS

Open-source cyber-agent training; harness evaluation awareness; self-evolution effects on measured cyber capability.

AI SAFETY BACKGROUND

FlowGuard: co-first work on in-generation safety detection for diffusion models. arXiv:2604.07879.

Unbridled Icarus: survey of image-input security risks in multimodal LLMs. arXiv:2404.05264.

Last updated: June 2026.