Changyi Li

changyi.li.0610@gmail.com | cosmosyi.github.io | CosmosYi | Google Scholar

EDUCATION

Fudan University & Shanghai Innovation Institute (SII) Sep 2026 - Jun 2030 (Expected)
Ph.D. in Artificial Intelligence *Shanghai, China*
– Advisors: Asst. Prof. Xudong Pan and Prof. Min Yang.

Fudan University, System Software and Security Lab Sep 2024 - Jun 2026 (Expected)
*M.S. in Computer Science; GPA: **3.96/4.00**, Rank: **1/379*** *Shanghai, China*
– Advisors: Asst. Prof. Xudong Pan and Prof. Min Yang.

Qilu University of Technology (Shandong Academy of Sciences) Sep 2020 - Jun 2024
*B.S. in Artificial Intelligence; GPA: **4.50/5.00**, Rank: **1/102*** *Shandong, China*

RESEARCH INTERESTS

My research focuses on **safety**, **alignment**, and **capability development** for **frontier AI systems**. I study reliability, trustworthiness, and controllability issues that arise as large models and agents tackle increasingly complex tasks and expand their capability boundaries. My current work focuses on scalable oversight and automated risk evaluation for advanced AI systems, as well as self-improving agent systems that learn from experience.

SELECTED HONORS AND AWARDS

First-Class Scholarship, Fudan University 2026
Freshman Scholarship, Fudan University 2025
Principal Scholarship ***(1/38,000+)*** 2024
Provincial Outstanding Graduate ***(Top 0.1%)*** 2024
National Scholarship ***(Top 0.1%)*** 2023
MCM/ICM Meritorious Winner, Team Leader 2023
First-Class Scholarship & Merit Student Pacesetter 2022, 2023

PUBLICATIONS

[1] AutoControl Arena: Synthesizing Executable Test Environments for Frontier AI Risk Evaluation.
Changyi Li, Pengfei Lu, Xudong Pan, Fazl Barez, Min Yang
ICML, 2026. [Paper] [Code] [Page] [Machine Heart] [PaperWeekly]

[2] ReasoningShield: Content Safety Detection over Reasoning Traces of Large Reasoning Models.
Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, Min Yang
Under review at NeurIPS, 2026. [Paper] [Code] [Models] [Data]

[3] CyberEvolver: Structured Self-Evolution for Cybersecurity Agents On the Fly.
Yihe Fan, **Changyi Li**, Lichen Xu, Xudong Pan, Jiarun Dai, Hong Geng, Min Yang
Under review at NeurIPS, 2026. [Paper] [Code] [Page]

[4] Large Language Model-Powered AI Systems Achieve Self-Replication with No Human Intervention.
Xudong Pan, Jiarun Dai, Yihe Fan, Minyuan Luo, **Changyi Li**, Min Yang
arXiv, 2025. [Paper] [Page] [LiveScience] [Forbes] [The Independent]

RESEARCH EXPERIENCE

AutoControl Arena: Automated Frontier AI Risk Evaluation Platform First Author | Jul 2025 – Mar 2026

- **Motivation:** Autonomous agents greatly expand the risk surface of AI systems, creating long-tail frontier risks; manual benchmarks are costly and hard to scale, while pure LLM simulators lack causal realism.
- **Method:** Built an Architect–Coder–Monitor framework: Architect structures natural-language risks, Coder synthesizes executable sandboxes via logic–narrative decoupling, and Monitor audits trajectory-level risks from behavior and reasoning traces.
- **Contribution:** Created AutoControl Arena for end-to-end automated risk evaluation; released X-Bench with 70 scenarios across 15 domains and 7 risk categories; introduced Stress–Temptation testing for long-tail frontier risks.
- **Results:** Achieved **98%** end-to-end construction success; reproduced OpenAI/Anthropic high-risk behaviors and identified alignment illusion, bidirectional safety scaling, and stress–temptation coupling effects.
- **Impact:** Received **100+ GitHub stars**, **600+ clones**, and media coverage from *Machine Heart* and *PaperWeekly*.

ReasoningShield: CoT Content Safety Auditing for Reasoning Models First Author | Feb 2025 – Jun 2025

- **Motivation:** CoT reasoning introduces hidden safety risks: harmful content can appear in long intermediate reasoning even when the final answer looks harmless or the model later refuses.
- **Method:** Built a 3-level, 10-category CoT risk taxonomy and trained **ReasoningShield-1B/3B** with SFT and DPO for harmfulness detection and step-level risk localization.
- **Contribution:** Introduced the **first CoT safety-auditing framework**, including a moderation benchmark with **9.2K** annotated query–CoT examples and a specialized moderation model for reasoning models.
- **Results: ReasoningShield-3B** reached SOTA on CoT safety benchmarks with **91.8%** F1, improving over strong baseline LlamaGuard-4-12B by **35.6%**; on mainstream QA safety benchmarks, it reached **85.0%** F1 (**+15.7%**).
- **Impact:** Released models and data with **2K+ Hugging Face downloads**; cited by teams including **IEEE Fellow**, **IAPR Fellow**, and **National Research Council Canada (NRC)**.

CyberEvolver: Self-Evolving Cybersecurity Agents Second Author | Feb 2026 – Jun 2026

- **Motivation:** Existing cyber agents rely on fixed human-designed workflows that struggle across targets and failure modes, while generic self-evolution faces unstructured mutations, sparse and noisy execution feedback, and limited update diversity.
- **Method:** Decomposed the cybersecurity-agent Harness into four structured evolvable layers; introduced trajectory diagnosis to turn noisy execution logs into actionable mutation signals; used population-based beam search to preserve diverse agent variants.
- **Contribution:** Proposed an on-policy Harness self-evolution framework for cybersecurity tasks; on CTF, vulnerability-exploitation, and penetration-testing tasks, improved seed-agent pass@16 by **13.6%** on average and outperformed the strongest human-designed baseline by **14.0%**.

AI System Self-Replication Risk Study Core Experimental Contributor | Nov 2024 – Feb 2025

- **Motivation:** Model self-replication is widely viewed as an AI safety red line, yet frontier AI labs at the time (e.g., OpenAI and Google) had underestimated the associated risk.
- **Method:** Designed a Perceive–Reflect–Discover–Plan harness to elicit longer-horizon self-replication beyond standard ReAct agents; evaluated 32 open-source models under a reproducible milestone-based protocol with explicit success criteria.
- **Contribution:** Provided the **first systematic empirical evidence** of end-to-end self-replication in open-source AI systems: **11/32 systems** succeeded, including a locally runnable **14B** model; further examined spontaneous replication, self-exfiltration, and shutdown evasion.
- **Impact:** Received **millions of views and reposts** on X, YouTube, and TikTok, coverage from *LiveScience*, *Forbes*, *The Guardian*, and *The Independent*, plus an invited presentation at a France AI Action Summit side event.

Text-to-Image Systems Red Teaming (Alibaba Tianchi Competition) Technical Lead | Sep 2024 – Nov 2024

- **Motivation:** Commercial text-to-image systems face full-chain content-safety and misuse risks, but many attacks target only a single model rather than the combined text-filter, generator, and image-filter stack.
- **Method:** Designed cognition-deception and semantic-association attacks with scenario rewriting, adversarial affixes, caption injection, prompt fuzzing, and VLM pre-screening.
- **Contribution:** Led full-chain gray-box and black-box vulnerability discovery across **30+** commercial systems; achieved a **90%+** text-filter bypass rate and **80%+** image attack success rate; ranked **26/1574** teams (**top 1.7%**); escalated severe vulnerabilities in **10+** mainstream commercial models to a **national security briefing**, which received central-level attention.

INDUSTRY PROJECTS

Full-Stack Multimodal Guardrail System Project Lead | Alibaba Cloud / CAS | Sep 2025 – Mar 2026

- **Background:** Enterprise multimodal review must cover fine-grained unsafe imagery, implicit semantic risks, and evolving compliance requirements, while real risk samples are scarce, human labels are costly, and existing pipelines are unreliable and hard to trace.
- **Method:** Built an automated data synthesis and labeling loop: constructed a risk lexicon and risk-category/concept space, expanded abstract risk terms into visual scene seeds and full image prompts, and used structured VLM review, multi-model voting, consistency scoring, and confidence calibration to produce traceable high-confidence pseudo-labels with hard-example feedback and incremental training.
- **Outcomes:** Achieved **95%+ evaluation accuracy** and **90%** efficiency improvement; supported multimodal content-safety evaluation and guardrail work for **Alibaba Cloud**, **CAS**, **China Electronics**, and **E Fund**.

AIGC Text Quality and Compliance Detection System Project Lead | China Electronics | Jun 2025 – Sep 2025

- **Background:** China Electronics (Fortune 500) needed production-grade quality acceptance and compliance screening for AIGC artistic-text outputs, including text omissions, malformed glyphs, semantic mismatch, and sensitive content; general VLMs reached only about 60% accuracy on fine-grained defects.
- **Method:** Designed a five-stage lightweight detection pipeline: OCR and rule checks for text omissions, VLM-based image-text consistency judging, multimodal fine-tuning, ViT defect classification, and model ensembling for joint quality and compliance decisions.
- **Contribution:** Designed the defect taxonomy, constructed negative samples, trained models, and deployed the service, building a production detector across **8 problem types** and **10+ visual styles**, including typos, malformed glyphs, semantic mismatch, risky text, deformation, and aesthetic defects.
- **Outcomes:** Exceeded **95% negative-sample detection**, met a **500 ms** latency requirement, handled **500K+/day** peak volume, and delivered **10M+** high-quality data assets.

ACADEMIC SERVICE

Invited Talk	Seoul Alignment Workshop, **FAR.AI**	2026
Invited Participant	Safety & Alignment closed-door mixer (ICML), **OpenAI**	2026
Organizer	AI Deception Workshop, **Fudan** & **SAIF** & **Concordia AI** [Link]	2025
Teaching Assistant	Foundation of Deep Learning, Fudan University	2025, 2026