王志华 Carey

15885844762 · wzh2657751462@gmail.com
个人主页：https://kirito14it.github.io/ · GitHub 主页：https://github.com/Kirito14IT

2023 级 | ★ 仓库 star 总数约 2.5k | ACM 集训队员（银奖） | 国家励志奖学金 *5

教育经历

复旦大学 复旦计算与智能创新学院 电子信息-计算机（**直硕**） 2027.09–2030.07

课题组：**FVL**、**TEAI**（负责人：**姜育刚副校长**） 导师：**马兴军** 2025.11–至今

研究兴趣：LLM　Coding Agent　Agentic Training　TrustWorthy AI | 英语 CET-6/CET-4：516/522

上海理工大学（本科生） 光电信息与计算机工程学院 **计算机科学与技术（院士创新班）　排名：3/85** 2023.09–2027.07

学术论文与知识产权

[1] **Zhihua Wang**, Nan Li, Ziyang Chen, Yunhao Feng, Songyang Wu. *ScamWeb: A Multimodal Benchmark for Grounded Understanding of Cyber-enabled Fraud Webpages*. Preprint; AAAI 2027（CCF-A）; Will be accepted(accept*4) **[First Author]**

[2] **Zhihua Wang**, Yunhao Feng, Yifeng Gao, Rui Pan, Baiheng Zhao, Zixing Chen, Xingjun Ma. *MemHarmBench: Evaluating Cross-Session Prompt Injection in Memory-Persistent AI Agents*. Preprint; WWW 2027（CCF-A） **[First Author]**

[3] **Zhihua Wang**, Yifeng Gao, Yunhao Feng, Rui Pan, Zhouyuan Ma, Baiheng Zhao, Zixing Chen, Yanping Li, Yizhang Liu, Xingjun Ma. *Smooth Inliers, Erratic Outliers: Frequency-Aware Mamba for Two-View Correspondence Learning*. Preprint; IEEE TIP（中科院一区 Top，CCF-A） **[First Author]**

[4] **Zhihua Wang**, Junfan Chen, Zixiang Wei, Lan Lin, Guoxiang Tong. *SafeCodeRL: Security-Constrained Multi-Agent Reinforcement Learning for Trustworthy LLM-Generated IoT/CPS Software*. Sensors, 26(11), 3502 Accept（CCF-C） **[First Author]**

[5] **Zhihua Wang**, Yifeng Gao, Yunhao Feng, Rui Pan, Baiheng Zhao, Zixing Chen, Nan Li. *CST: Lightweight Risk Triage with Source-Grounded Evidence for Web Discussions*. Preprint; arXiv:2611.01710 **[First Author]**

[6] Yifeng Gao, Yifan Ding, **Zhihua Wang**, Zixing Chen, Rui Pan, Xin Wang, Yixu Wang, Ye Sun, Yutao Wu, Hanxun Huang, Xingjun Ma. *HorizonGuard: Runtime Risk Forecasting for LLM Agents*. Under review; ICLR 2027 **[Third Author]**

软件著作权：《DL 驱动的鲁棒型数字水印网络应用》《基于 CodeQL 与 LLM 的代码安全分析系统》**第一作者**

相关专利：信息网络安全公安部重点实验室（公安部第三研究所）相关专利申请正在评审。

研究经历

SafeCodeRL：面向可信代码生成的安全约束多智能体强化学习框架

2026.02–2026.05

Multi-Agent Code Generation for SFT & RL | PPO + Lagrangian CMDP | CodeQL Audit + Docker sandbox | vLLM + SWE-bench

- **背景与目标：**IoT/CPS 软件越来越多由 LLM/Agent 生成，SQL 注入、路径穿越、命令注入等漏洞会危及感知数据与系统安全；现有方法多为事后静态过滤，难以在功能与安全之间做有原则的权衡。目标：**在生成过程中显式施加安全约束，保持功能正确率的同时大幅降低高危漏洞。**
- **我的职责：项目 Owner**（第一作者），从 0 到 1 负责问题建模、框架设计、训练与评测实现及论文写作：
  - **Agent 架构：**将代码生成建模为 **CMDP**，设计 Planner/Code/Security/Test/Critic **五智能体闭环**，覆盖**任务规划**、代码生成与**自我修正**、漏洞审计和 **Docker sandbox 测试**（10 s 超时、512 MB、断网），由 Critic 价值网络与路由 MLP 动态调度修复/重测/终止。
  - **安全审计：**Security Agent 结合 **Python AST** 检查与 **CodeQL 数据流（污点）分析**定位 CWE-89/22/78 等漏洞，再以规则复核过滤误报，输出带严重度的结构化漏洞元组作为安全代价。
  - **后训练：SFT** warm-start 后以 **PPO + Lagrangian 自适应安全惩罚**优化 Code Agent，AST 级 shielding 拦截高危代码；设计“编译 → 运行 → 测试通过率 → Pass@k”**渐进式稠密奖励**缓解奖励稀疏；基于 AutoGen 与 **vLLM** 张量并行在 8×A100 上完成多智能体 rollout 与训练。
- **实验结果：**GPT-4o 底座下较 AgentCoder，**SWE-bench** 解决率 13.1% → **14.8%**、HumanEval Pass@1 达 **91.4%**，SecurityEval/CyberSecEval 漏洞率 17.6%/16.4% → **5.5%/6.2%**（降幅超 **60%**）；25 个 IoT/CPS 任务 Secure Pass@1 达 **89.3%**；消融中去掉 Lagrangian 约束漏洞率升至 17.5%，Qwen2.5-Coder 版较 DeepSeek-R1 少用约 **70%** token。

论文：https://kirito14it.github.io/papers/safecoderl.pdf

ScamWeb：面向网络诈骗网页检测、细分类与证据定位的多模态基准

2026.04–2026.08

Training-free LLM-VLM Cascade | Playwright + Wayback Machine + MHTML Audit | GPT-4o-mini + Qwen3-VL-Flash

复旦大学可信具身智能研究院**马兴军** & **姜育刚**老师实习学生 | AAAI 2027（accept*4） | **第一作者**

- **背景：**网络诈骗已成为全球性的公共安全问题，但现有公开数据集多聚焦钓鱼与品牌仿冒、只有粗粒度标签，缺少覆盖多类型、多语言诈骗网页，并支持证据定位与可解释研判的基准。目标：**构建高质量多模态基准，并给出无需按任务训练的统一基线。**
- **我的职责：项目 Owner**（第一作者），负责数据整合与清洗、任务与指标设计、ScamDet 框架实现、全量评测及论文写作：
  - **数据管线：**汇聚官方黑名单、社区举报、安全厂商情报等约 **10 万**条候选 URL，经 **Playwright** 在线爬取与 **Wayback Machine** 回溯复原、整页截图与 MHTML 存档、pHash 一致性审计，清洗为 **10,026** 页、**25** 类诈骗、**47** 种语言的基准，其中 **3,416** 页含 **5,489** 个专家证据框及解释。
  - **统一任务与框架：**定义检测–细分类–证据定位–解释生成统一任务 **GroundScam**；提出免训练 **ScamDet**，以 GPT-4o-mini 分析 URL 与文本、Qwen3-VL-Flash 分析视觉，**四阶段级联推理**，每阶段**严格 schema 输出**（有效输出率 **100%**）。
- **实验结果：**与 GPT-4o、GPT-5 mini、Claude Sonnet 4、Gemini 3 Flash、Qwen3-VL-235B 等 **8 个前沿模型零样本**对比，在 **3,324** 页诈骗子集上召回率 **97.3%**、细分类准确率 **63.9%**、定位 mIoU **0.193** 均列第一；二分类 F1 达 **78.7%**。

论文：https://kirito14it.github.io/papers/scamweb.pdf **代码：**https://github.com/Kirito14IT/ScamWeb **数据：**https://huggingface.co/datasets/kirco567/ScamWeb-Dataset
项目网站：https://kirito14it.github.io/ScamWeb/

VulnSeeker：基于 CodeQL 与大模型的代码安全分析系统

2026.01–2026.03

CodeQL Static Analysis + LLM Review | Call-Chain Context Retrieval | FastAPI + Socket.IO + React + MySQL

复旦大学软件工程实验室（CodeWisdom） | 指导：**李弋** | 已登记软件著作权

- **背景与目标：**CodeQL 等静态分析在跨函数调用链下误报高，告警需要大量人工复核；原型工具 Vulnhalla 只提供 CLI/TUI 交互，难以支撑多任务、多仓库的审查流程。目标：**用 LLM 对静态告警做上下文感知的二次复核，并沉淀为可多人使用的 Web 审查平台。**
- **我的职责：项目 Owner**，基于 CyberArk Vulnhalla 原型二次开发，负责分析引擎改造与平台前后端：
  - **程序分析 + LLM：CodeQL** 静态分析定位候选漏洞，LLM 按需请求**外部函数与调用链**上下文做**二次复核与分类**，以状态码输出研判结论。
  - **平台化：**重构为前后端分离 Web 平台，支持 GitHub/本地 **CodeQL 数据库**与源码构建三种任务模式、**任务级隔离工作区**与 Socket.IO 实时进度，统一展示 CodeQL 告警与 LLM 复核结果，并支持 TP/FP/Uncertain **人工标注**。

代码：https://github.com/Kirito14IT/VulnSeeker

MemHarmBench：面向持久记忆投毒的跨会话 LLM Agent 安全基准 上海创智学院 Shanghai Innovation Institute 2026.07–2026.10

复旦大学计算与智能创新学院**姜育刚** & **马兴军**导师访问学生 上海创智学院推荐论文 奇绩创坛 Alpha Sight 推荐论文 | **WWW 2027** | **第一作者**

- **研究动机：**长期记忆让 Claude Code、OpenClaw 等 **Coding Agent** 跨会话复用规则，也让低权限协作者或外部内容可植入持久恶意规则，并在后续无攻击载荷的任务中触发；单轮注入评测无法刻画这类风险。目标：**构建跨会话记忆投毒基准与评测协议，量化“被操纵却看似正常”的失败**。
- **我的职责：项目 Owner**（第一作者），主导威胁模型、基准协议与指标设计，负责数据集、评测 harness、全部实验与论文写作：
  - **评测协议：**提出“**暴露–触发**”双会话协议：先检验恶意规则是否写入记忆，再移除载体、在新会话执行任务，将跨会话影响与重复暴露解耦。
  - **基准与 harness：**按 3 类受害者 × 12 类攻击目标设计威胁分类，覆盖**代码仓库维护**、**CI/CD 与云运维**、RAG 检索等 12 个业务场景，共 **288** 个可复现案例；开源 pip 可安装的**评测 harness** 与 CLI，以固定版本 **Docker 镜像**原生运行 Agent，GPT-5 按证据准则做 LLM-as-Judge。
- **实验结果：**在 OpenClaw/Claude Code × GPT-5/Gemini-2.5-Flash 四种配置上，跨会话间接攻击成功率 **59.38%–76.39%**，且 **55.71%** 的回合在交付完整、轨迹无异常的同时已被操纵（“**隐蔽寄生**”）；受控网页实验中，**视觉提示注入**使采纳率由 2.5% 升至 **40.0%**。

论文：https://arxiv.org/abs/2611.01700 **代码：**https://github.com/Kirito14IT/MemHarmBench **数据：**https://huggingface.co/datasets/kirco567/MemHarmBench
项目网站：https://kirito14it.github.io/MemHarmBench/

SFMambaNet：谱频增强选择性状态空间对应剪枝模型 2025.09–2026.02

Frequency-Aware Mamba | **FFT Gating + Spectral-Geometric Attention** | **PyTorch + mamba-ssm** | **YFCC100M / SUN3D**

福州大学福建省网络计算与智能信息处理重点实验室 | 导师：刘翼章 | IEEE TIP（中科院一区 Top，CCF-A） | **第一作者**

- **技术挑战：**双视图对应剪枝需要在大量外点中识别内点，基于 Transformer 的方法全局注意力为二次复杂度，且没有利用内外点在频域上的差异。目标：**以近线性复杂度建模全局上下文，并稳健抑制外点**。
- **我的职责：项目 Owner**（第一作者），主导选题、模型设计、代码实现、训练调参、消融与对比实验及论文写作：
  - **核心洞察与模型：**发现匹配点按聚类顺序扫描时内点呈**低频平滑**、外点呈**高频跳变**，据此设计**局部谱几何注意力**与**频域门控双向 Mamba**，以**近线性复杂度**（$O(N \log N)$）抑制高频外点噪声。
- **实验结果：**YFCC100M 未知场景（SIFT，无 RANSAC）mAP5° 达 **73.83**，较同协议复现的 MatchMamba 提升 **6.23** 个百分点，参数量仅 **2.01M**；**五随机种子对照实验**表明增益来自频域滤波而非参数容量。

论文：https://arxiv.org/abs/2606.04493 **代码：**https://github.com/Kirito14IT/SFMambaNet **项目网站：**https://kirito14it.github.io/SFMambaNet/

XSafeClaw：Agent 运行时安全治理的可信防护框架 ★ Stars 165 获「机器之心」报道 2026.03–2026.06

Agent Runtime Guardrails | **AgentDoG + before_tool_call Hooks + Human Approval** | **FastAPI + WebSocket + React**

复旦大学可信具身智能研究院 | 导师**马兴军** | 登上可信具身智能研究院首页“最新研究”

- **背景与目标：**能执行代码、调用工具、访问本地资产的 Agent 缺少统一的运行时安全接入与风险治理，OpenClaw、Hermes、nanobot 等异构运行时难以被统一发现、监控与保护。目标：**构建跨运行时的“可观测–可拦截–可审批–可审计”安全闭环**。
- **我的职责：Core Contributor**，主导 nanobot 运行时端到端接入，并参与守卫链路、多运行时统一与版本迭代：
  - **双层守卫：AgentDoG** 对会话轨迹做**多轮风险评估**，**before_tool_call 钩子**逐次拦截**工具调用**，高风险操作挂起转**人工审批**，联动资产扫描。
  - **nanobot 接入：**完成 Runtime Registry/Gateway 会话发现、**Guard Hook** 注入与配置治理、安全模板逐轮注入、Safe Chat 与 Agent Valley 统一展示，统一**异构运行时**的会话身份与工具事件。
  - **工程落地：**实现**并发多会话通信**与三类运行时按模型 **Token/成本统一核算**，完善运行时防护与**集成测试**；参与 v1.0–v1.1 版本迭代与 PyPI 发布。

项目网站：https://xsafeclaw.ai **代码：**https://github.com/XSafeAI/XSafeClaw **报道：**https://mp.weixin.qq.com/s/1Dp5uTasxnm3ues_hNPlWQ

实习经历

公安部第三研究所 信息网络安全公安部重点实验室 (公安部第三研究所上海) AI 研究实习生 2026.04–至 今

- **背景与方向：**面向反诈与网络内容治理一线需求，导师吴松洋；负责多模态诈骗网页研判与群聊风险筛查两条算法线。

项目 1：ScamWeb 诈骗网页多模态研判

- **我的职责：项目 Owner**，独立构建网页图文与 URL 联合分析及诈骗证据定位流程；完成**独立一作论文 1 篇**（AAAI 2027，详见研究经历）。

项目 2：CST 群聊短文本风险分诊（arXiv:2611.01710）

Lightweight Risk Triage | **MiniLM + Char n-gram + Logistic Regression** | **Hierarchical Clustering** | **CPU Deployment**

- **背景与目标：**群聊与讨论平台消息量大、语言混杂，审核人员需要知道看哪些会话、读哪些消息、怎样回到原文核对，且只能在本地 CPU 上运行。目标：**轻量风险分诊，并给出可回链原文的证据**。
- **我的职责：项目 Owner**（第一作者），从 0 到 1 设计算法、实现本地应用并完成论文：
  - **算法：**设计**轻量级短文本风险主题发现算法**：冻结的多语言 MiniLM 句向量（384 维）与字符 2–5 gram 特征接三个逻辑回归头，识别消息风险并排序会话；余弦层次聚类与原文短语抽取组织主题证据，每条证据可回链原始消息与回复路径。
  - **本地应用：**React + Express 串行队列 + 常驻 Python worker + SQLite，支持原文定位、审阅笔记与 CSV/JSON 导出；CI 覆盖测试与前端构建。
- **实验结果：**6 个消息级风险任务 AP 均超过全量微调的 XLM-R，CPU 处理速度为其 **2.6** 倍；人工审核实验中，分组证据使审核时间减少 **27.3%** 以上、判定准确率不低于 **94.7%**；论文已发布 arXiv（第一作者），相关专利申请评审中。

国际组织（联合国志愿人员组织）–UNDP **数据研究实习生（线上）** 远程负责团队数据的清洗、分析与可视化。 2026.01–2026.04

竞赛经历

2025 ACM 上海站大学生程序设计竞赛	**银 奖**	2025.07
2025 年第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛 C/C++	**国家级二等奖**	2025.06
2025 年睿抗机器人开发者大赛编程赛道全国总决赛	**国家级一等奖**	2025.08

技能与服务

编程与工程： C/C++、Python；Linux/Shell、Git；Docker sandbox、FastAPI、自动化与集成测试
大模型与 Agent： PyTorch/TensorFlow、Transformers、vLLM；SFT、PPO/约束强化学习；AutoGen、Codex、Claude Code
评测与程序分析： SWE-bench、HumanEval、SecurityEval、CyberSecEval；CodeQL 静态分析；**NeurIPS2026 审稿人**

个人总结

- **科研产出与学术服务：**本科期间研究经历多段，研究时间跨度超过 1 年，多篇第一作者论文（AAAI 等），**NeurIPS 2026 审稿人**，MemHarmBench 获两家机构推荐。
- **开源影响力：**GitHub 仓库 star 总数约 **2.5k**；XSafeClaw 核心贡献者（**165** Stars、机器之心报道）；MemHarmBench、ScamWeb 数据与代码均已开源。
- **AI 原生工作流：**日常以 **Claude Code / Codex / Gemini 4 Carbon** 驱动科研实验、数据合成与评测、代码开发及 Agent harness 搭建。
- **算法竞赛功底：**ACM 集训队员（上海站**银奖**）、蓝桥杯全国总决赛 C/C++ **国家级二等奖**、睿抗机器人开发者大赛全国总决赛**国家级一等奖**。