Reworr

Address: Cracow, Poland
Email address: reworr@protonmail.com
Web: www.linkedin.com/in/reworr

Profile

AI Security Researcher with a background in offensive cybersecurity.

I evaluate and red-team frontier AI models, focusing on autonomous hacking capabilities, spear-phishing, and loss-of-control scenarios. My work has been covered by MIT Technology Review, Bloomberg Law, Cybernews, and TIME.

Before moving into AI security, I spent several years in penetration testing and vulnerability research. I have disclosed vulnerabilities credited by Oracle, Telegram, and Meta Research. I have been an active CTF player since 2016 and was a member of a top-30 worldwide team.

Experience

2026
Research Scholar (UK AISI Red-Team)
ML Alignment & Theory Scholars (MATS)
AI red-teaming research within the UK AI Security Institute (AISI) Red Team.

2024 – 2026
AI Security Researcher
Palisade Research
Evaluations of LLMs' offensive capabilities (e.g., autonomous hacking, spear-phishing, large-scale OSINT, Loss of Control).
Preparing technical demos for Palisade's briefings to policymakers and government agencies.
Red-teaming frontier LLMs within pre-release access programs.

2025
LLM Red Teaming Contractor
Trajectory Labs
Designed novel prompt-injection eval scenarios and adversarial variants for a frontier AI lab (NDA).

2024
Research Fellow
Apart Research
Lead author of the "LLM Agent Honeypot" project (https://www.apartresearch.com/post/hunting-for-ai-hackers-in-the-wild-llm-agent-honeypot).

Experience

2023 – 2024
Penetration tester
Deteact
Web/Mobile Application Security Analysis.
Social engineering, Spear-phishing attacks.

2023
Security Analyst
CleanTalk Inc
Conducted web security/vulnerability research.
Performed forensics and remediation/hardening.

2022
Web Security Fellow
eQualitie
Audited web apps and infrastructure for NGOs in high-risk contexts; guided remediation efforts.

2021 – 2022
Penetration tester
Engineering Center Regional Systems
White/Black-box penetration testing (web-security testing, internal pentest, phishing).

Publications

LLM Agent Honeypot: Monitoring AI Hacking Agents in the Wild
https://arxiv.org/abs/2410.13919
An open-source honeypot project featured by Bloomberg Law, MIT Technology Review, Cybernews.

AI Hacking Cable (PoC/demo)
https://palisaderesearch.org/blog/hacking-cable
An autonomous LLM agent for post-exploitation operations. Mentioned in TIME magazine (https://time.com/7316051/chatbots-parroting-russian-propaganda/).

GPT-5 at top-tier CTFs
https://arxiv.org/abs/2511.04860
We evaluated GPT-5's cybersecurity capabilities by entering it in top-tier CTF competitions, where it placed 25th in one of the year's hardest events, outperforming 93% of human teams.

Publications

The frontier of AI Security: 2024
AI Security Newsletter
https://www.heronsec.ai/post/the-frontier-of-ai-security-what-did-we-learn-in-the-last-year
Authored year-in-review analysis of AI security challenges and breakthroughs, covering jailbreak vulnerabilities, AI-enabled cyber operations, model security, and emerging defenses.

Evaluating AI cyber capabilities with crowdsourced elicitation
https://arxiv.org/abs/2505.19915 (Ack. contributor)
Ran an LLM-based CTF agent in the "AI vs. Humans" cybersecurity competition (Hack The Box, 400 teams), solving 19/20 challenges and placing 2nd among AI teams.

Public Talk: "Offensive Use of LLMs: Current Capabilities & Risks"
https://reworr.com/bsides-krakow-2025
Talk at BSides conference on AI capabilities in offensive security, featuring projects I worked on.

Public Talk at AI Safety Poland
https://reworr.com/ai-safety-poland-2026
Talk covering LLM offensive capabilities from autonomous hacking to rogue replication, featuring projects I worked on.

Awards

Runner-up, Technical Research Prize
Bluedot Impact AI
https://blog.bluedot.org/p/llm-agent-honeypot-monitoring-ai-hacking-agents-in-the-wild
Awarded for the LLM Agent Honeypot project in BlueDot Impact's AI Alignment course.