Reworr R.

Cracow, Poland reworr@protonmail.com www.linkedin.com/in/reworr

Profile

AI Security/Safety Researcher with a background in Offensive Cybersecurity.

Current focus:
- AI Security & AI Red Teaming (e.g., adversarial attacks)
- AI Safety & AI Capabilities Research (mostly in Cybersecurity)

Cybersecurity experience:
- Penetration Testing

External & internal pentests, social-engineering/spear-phishing campaigns.
- Vulnerability Research

Credited by Oracle and Telegram for disclosed vulnerabilities, as well as by open-source projects (e.g., *CVE-2022-25876*).
- CTF player since 2016; ex-member of "MindCrafters", a top-30 team worldwide.

Work Experience

2024 – PRESENT
AI Security Researcher
Palisade Research

Examples of projects & responsibilities:
- Conduct evaluations of LLMs' offensive capabilities (e.g., autonomous pentesting, spear-phishing, large-scale OSINT).
- Develop and evaluate jailbreak method.
- Lead red-teaming of frontier LLMs within early-access/safety programs (e.g., OpenAI, Anthropic).

2024
Research Fellow
Apart Research at Apart Lab

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

2023 – 2024
Penetration tester
Deteact

Web/Mobile Application Security Analysis.
Social engineering, Spear-phishing attacks.

Work Experience

2023
Security Analyst
CleanTalk Inc

Conducted web security research and investigations. 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 MIT Technology Review, Forbes, Cybernews.

Public Talk: "AI in Offensive Security"
https://reworr.com/bsides-krakow-2025
Gave a talk at the BSides conference on AI capabilities in offsec, featuring projects I worked on.

Evaluating AI Capabilities in Cybersecurity: GPT-5 at Top-Tier CTFs
https://github.com/PalisadeResearch/gpt5-ctfs/releases/download/latest/main.pdf

AI Malware PoC
https://palisaderesearch.org/assets/reports/hacking-cable-report.pdf
https://x.com/i/status/1963596598728110588

The frontier of AI Security: 2024
AI Security Newsletter

https://securityai.substack.com/p/the-frontier-of-ai-security-what