

Reworr R.

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



👤 Profile

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

Current focus:

- AI Red Teaming (e.g., adversarial attacks)
- Offensive AI capabilities (mostly 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).
- Lead red-teaming of frontier LLMs within early-access programs (e.g., OpenAI, Anthropic).

⌚ 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 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/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

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.

GPT-5 at CTFs: Case Studies From Top-Tier Cybersecurity Events

<https://arxiv.org/abs/2511.04860>

Conference Talk: "AI in Offensive Security: Capabilities & Trends"

<https://reworr.com/bsides-krakow-2025>

Talk at BSides conference on AI capabilities in offensive security, featuring projects I worked on.

AI Malware (PoC)

<https://palisaderesearch.org/blog/hacking-cable>

Built an autonomous LLM agent for post-exploitation operations.

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>

Evaluating AI cyber capabilities

<https://arxiv.org/abs/2505.19915> (Ack. contributor)

Run a Claude-based agent; 2nd among AI teams.