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:
- Offensive AI capabilities (mostly cybersecurity)
- AI Red Teaming (e.g., adversarial attacks)

Cybersecurity experience:
- Penetration Testing
External & internal pentests, social-engineering/spear-phishing campaigns.
- Vulnerability research
Credited by Oracle and Telegram for disclosed vulnerabilities, by Meta Research for a bypass in Meta-SecAlign (SOTA AI Defense), as well as by open-source projects (e.g., *CVE-2022-25876*).
- CTF player since 2016; ex-top-30 worldwide team member ("MindCrafters").

💼 Work Experience

📅 2024 – PRESENT
AI Security Researcher
Palisade Research
Evaluations and demos of LLMs' offensive capabilities (e.g., autonomous hacking, spear-phishing, large-scale OSINT, Loss of Control).
Red-teaming frontier LLMs within pre-release 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
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.

AI Malware (PoC/demo)
https://palisaderesearch.org/blog/hacking-cable
An autonomous LLM agent for post-exploitation operations. Mentioned in TIME magazine.

GPT-5 at top-tier CTFs
https://arxiv.org/abs/2511.04860
Case Studies From Top Cybersecurity Events

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.

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.