Jonas Wiedermann-Möller

Bielefeld, Germany | contact@jowimo.com

RESEARCH EXPERIENCE

Exponential Security Labs GmbH (ExpSec) *2026*
Research Intern, AI Security & Agentic Systems | Tübingen

- Contributed to RepoInject v1 and BrowserInject v1, prompt-injection benchmarks for coding and browser agents used in the evaluation of Meta's Muse Spark 1.3.
- Built an automated pipeline for generating reinforcement-learning environments for agentic systems.
- Built reproducible evaluation and verification workflows for agentic security tasks and scalable experimentation.

EDUCATION

Bielefeld University Bielefeld
Master of Science in Data Science, Final Grade: 1.7 *2022 – 2026*

- Thesis (**1.5**): *Instrumental Choices*; supervised by Maksym Andriushchenko at ELLIS Institute Tübingen
- Erasmus stay: Université Gustave Eiffel, France (2024–2025)

Fachhochschule des Mittelstands Bielefeld
Bachelor of Arts in Digital Business Management, Grade: 1.7 *2018 – 2021*

- Thesis on the opportunities and limitations of low-code development platforms for business applications

RESEARCH & PROJECTS

Instrumental Choices: Measuring the Propensity of LLM Agents to Pursue Instrumental Behaviors
Lead Author | Accepted at NeurIPS 2026 2026
Developed an evaluation framework measuring resource acquisition, shutdown resistance, and persistence in LLM agents. Presented to the red-teaming group at the UK AISI. arxiv.org/abs/2605.06490

Agent Incidents 2026
Built and maintain a sourced public timeline of reported AI agent incidents, distinguishing observations from interpretations and training, evaluation, and deployment settings. agent-incidents.com

Independent investigations into AI agent activity 2026
Investigated public-web traces of AI agents, with findings included in the Collusion Wiki updates. This work received coverage in The Washington Post, Reuters, DER SPIEGEL, and Axios.

Open source & writing
Created markitdown-gui, a GUI for Markdown conversion. Write at jowimo.com/writings; published "How I Use Codex to Automate Parts of My Research" for the AISA Group.

SKILLS & PERSONAL

Language: German (native), English (C2), French (B1)
Research: AI safety evaluation, adversarial red-teaming, LLM agent evaluation, RL environment design, synthetic data and environment generation, evaluation engineering, reproducible experiments
Stack: Python, PyTorch, Inspect AI, Hugging Face Transformers, Prime Intellect (verifiers, prime-rl), Docker, Slurm, Git/GitHub, OpenAI Codex

OTHER EXPERIENCE

pytchBlue OÜ *2021 – 2025*
Co-Founder | IT Support and Consulting | Bielefeld, Tallinn

- Delivered analytics, infrastructure optimisation, and cloud automation for SMEs.

Earlier experience (2019 – 2022): ALDI International Services SE (Senior IT Professional), KPMG AG (Associate Consultant), and mindsquare AG (Working Student & Intern), spanning SAP infrastructure integration, authorisation, compliance, and security automation.