Thomas Winninger

M2 mathematics student at Télécom SudParis - ENS Paris-Saclay

Mail: thomas [dot] winninger [at] telecom-sudparis [dot] eu **- LinkedIN:** thomas-winninger
Website: https://le-magicien-quantique.github.io **- GitHub:** Sckathach

Education

- 2025 - 2026 **Master MVA, ENS Paris-Saclay**
  Topology, optimal transport, reinforcement learning, training and deploying large-scale models, LLM, graph neural networks, learning for protein science, convex optimization.
- 2022 - 2026 **Engineering Degree, Télécom SudParis**
  Telecommunications, cyber security, cloud, information theory, probability, optimization, graph theory, graph neural networks, signal processing.

Experience

- Sep 2025 - *now* - **Teaching and research sprints - PIAF**
  Teaching (interpretability, LLM training and fine-tuning) and organizing short research sprints (teams of 5 people, lasting under four days).
- Jul - Sep 2025 - **Research internship in LLM security - NICT**
  Research on the security and jailbreak interpretability of Large Reasoning Models (LRMs). *I studied LRM robustness, adapted state-of-the-art black-box and white-box attack from LLMs, and started studying jailbreaks with interpretability methods on LRMs.*
- Mar - May 2025 - **Research internship in AI explanability - INRIA**
  Verified robust explanation for language models. *I explored scaling Hybrid Constrained Zonotopes (HCZs) to language models using convex relaxation and optimization. However, the relaxation error proved too large for practical use.*
- Jul - Dec 2024 - **Research internship in AI security - Thales**
  Implementations and improvements of state-of-the-art attacks on LLMs. *I improved state-of-the-art white-box adversarial attacks on LLMs and published the results on ArXiv.*
- 2022 - 2024 - **Teaching and infrastructure - HackademINT**
  Teaching (cloud and AI security), cloud management (Kubernetes), creation of challenges (AI & quantum physics), and organization of 404CTF 2023 & 2024 (largest cyber security competition in France).

Miscellaneous

- Languages: **Python, French, OCaml, English**, Typst, TypeScript, Lua, Rust, C, Bash, Japanese (JLPT 4), Lean
- Tools/ Frameworks: **PyTorch, nnsight**, Docker (Podman), Kubernetes, React, Qiskit, Archlinux
- Other interests: Piano, guitar, teaching, reading, geopolitics, particle physics, sports, video game (playing & development), meditation
- I completed the Alignment Research Engineer Accelerator (ARENA) and the AI Safety Fundamental (AISF) curriculums.

Papers

- Scaling Hybrid Constrined Zonotopes with optimisation - *Winninger T., Urban C., Wei G., Jun 25*. Paper
- Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models - *Winninger T., Addad B., Kapusta K., Mar 25*. ArXiv / Webpage

Talks

- Adversarial attacks against reasoning LLMs, *Tokyo, NICT, Sep 25*.
- Scaling abstract domains to Large Language Models with Hybrid Constrained Zonotopes, *ENS Ulm, INRIA, Jun 25*.
- Mechanistic interpretability for LLM attack and defense, *École Polytechnique, CeSIA, Apr 25*. Slides
- Introduction to AI security and reverse engineering, *Télécom SudParis, HackademINT, Apr 25*. Slides / Webpage
- Model Poisoning, *Station F, CeSIA, Jun 24*. Slides
- GNN based IDS and its robustness against adversarial attacks, *Télécom SudParis, HackademINT, Jun 24*. Slides
- Cheating Detection in the 404 CTF, *Rendez-vous de la Recherche et de l'Enseignement de la Sécurité des Systèmes d'Information (RESSI), May 24*.
- Introduction to prompt hacking, *Télécom SudParis, HackademINT, Nov 23*. Slides
- How to backdoor federated learning, *Télécom SudParis, HackademINT, May 23*. Slides
- Introduction to AI & cyber security, *Télécom SudParis, HackademINT, May 23*. Slides

Research reports

- Graph Neural Network based Intrusion Detection and its Robustness against Adversarial Attacks, *Moreau R., Winninger T., Blanc G., Jun 24*. Paper

Hackathons

- ZaMark: Intellectual Property protection with Homomorphic Watermarking, *Privacy Preserving Hackathon, Zama, Sep 24, (finished 2nd)*. Slides