

# Thomas Winninger

## Étudiant à Télécom SudParis

→ *Version française ici:*

**Mail:** [thomas.winninger@telecom-sudparis.eu](mailto:thomas.winninger@telecom-sudparis.eu) - **Website:** <https://le-magicien-quantique.github.io> - **GitHub:** [Sckathach](#) - **LinkedIn:** [thomas-winninger](#)

## Whoami?

Aka *the quantum warlock, the masked camel, the fanOfThermodynamics, the whale orchestra conductor*, or just **Sckathach**. I'm a french student at Télécom SudParis, and soon, an AI researcher!

Fond of mathematics and physics, I ended up at Télécom SudParis where I focused on cyber security. As I quickly became interested in AI security, I decided to take a gap year to bring myself up to speed on the subject: AI security research, interpretability, tools, statistics; and since that's what I like best, I'm now doing a master's degree in fundamental AI, and I plan to continue with a thesis, most likely in the same field.

## Education

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

## Experience

- Jul - Sep 2025 - **Research internship in LLM security - NICT**  
Security and jailbreak interpretability on reasoning LLMs.
- Mar - May 2025 - **Research internship in AI explainability - INRIA**  
Verified robust explanation for language models.
- Jul - Dec 2024 - **Research internship in AI security - Thales**  
Implementations and improvements of state-of-the-art attacks on LLMs.
- 2022 - 2024 - **Training and infrastructure - HackademINT**  
Teaching (cloud and AI security), cloud management (Kubernetes), creation of challenges (AI & quantum physics), and organization of 404CTF 2023 & 2024.

## Misc

- 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 :), (see my [GitHub](#) for more)

- Other interests: Piano, guitar, teaching, reading, geopolitics, particle physics :), sports, video game (playing & development), meditation
- Followed ARENA and AISF

## Papers

- Scaling Hybrid Constrained Zonotopes with optimisation - *Winner T, Urban C., Wei G., Jun 25.* [Paper](#)
- Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models - *Winner 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., Winner T., Blanc G., Jun 24.* [Paper](#)

## Posts

- Subspace Rerouting: Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models. [Post](#)
- Exploring the use of Mechanistic Interpretability to Craft Adversarial Attacks. [Post](#)

## Hackathons

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