

Thomas Winneringer

M2 mathematics student at Télécom SudParis - ENS Paris-Saclay.

I'm looking for an internship starting April 2026, on LLM alignment, AI security, or more generally fundamental AI that would lead to a PhD.

Mail: thomas.winneringer@telecom-sudparis.eu - **Website:** <https://le-magicien-quantique.github.io>

LinkedIn: [thomas-winneringer](#)

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

- Jul - Sep 2025 - **Research internship in LLM security - NICT**
Research on the security and jailbreak interpretability of Large Reasoning Models (LRMs).
- 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 - **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

- Spoken languages: **French (native)**, **English (TOEIC)**, Japanese (JLPT 4)
- Programming languages: **Python**, **OCaml**, Typst, TypeScript, Lua, Rust, C, Bash, Lean

Papers

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

Selected Talks

- 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](#)