

Thomas Winninger

Étudiant à Télécom SudParis en année de césure

✉ thomas.winninger@telecom-sudparis.eu 🏠 <https://le-magicien-quantique.github.io> 🗣️ Sckathach 🌐 Thomas Winninger
📠 0009-0000-2783-3086

Rapidement

Fan de mathématiques et de physique, j'ai commencé un cursus de cybersécurité à Télécom SudParis et me suis très vite intéressé à la sécurité de l'IA. J'ai donc décidé de prendre une année de césure pour me mettre à niveau sur le sujet : recherche en sécurité de l'IA, interprétabilité, outils, statistiques; et comme c'est ce qui me plaît le plus, je compte continuer avec un master et une thèse, très certainement dans le même domaine.

Formation

Master en informatique - Cyber Sécurité

Télécom SudParis - Institut polytechnique de Paris (IPP)

2024 - 2026

Diplôme d'ingénieur - Spécialisation Cyber

Télécom SudParis

2022 - 2026

Télécommunications, sécurité des réseaux et applications web, théorie des graphes (application à l'IA et à la 6G). Théorie de l'informatique et bases de données. Traitement du signal et probabilités.

Expériences

Stage de recherche en explicabilité des modèles de langage

INRIA - ANTIQUE

mars - mai 2025

Explicabilité de modèles de langage par interprétation abstraite.

Stage de recherche en sécurité de l'IA

Thales - ThereSIS

juillet - décembre 2024

Implémentations et améliorations des attaques état de l'art sur les LLM.

Responsable formation / infrastructure

HackademINT

2023 - 2024

Création de challenges (IA & physique quantique), et organisation du 404CTF 2023 & 2024.

Présentations

- **Model Poisoning** - *AI Safety Meetup / Centre pour la sécurité de l'IA* (juin 2024)
- **Détection de la triche dans le 404 CTF** - *Rendez-vous de la Recherche et de l'Enseignement de la Sécurité des Systèmes d'Information* (mai 2024)

Papiers

- **Using Mechanistic Interpretability to craft Adversarial Attacks against Large Language Models** - *Winner T., Addad B., Kapusta K.* (mars 2025)

Compétences

Langages Python, Ocaml, TypeScript, Typst, Rust, Lua, C, Bash

Langages parlés Français, Anglais, Coréen, Japonais

Outils PyTorch, PyG, Docker (Podman), Kubernetes, React, Qiskit, Sage, Archlinux :)

Autres intérêts

Piano, guitare, création de jeux vidéo, lecture, géopolitique, physique des particules, sport, méditation, enseignement.