Oudoum **Ali Houmed**

M.Sc. Data, Knowledge and Hybrid Artificial Intelligence (DKAI) | Paris-Saclay University

Paris, France

oudoum.ali-houmed@universite-paris-saclay.fr oudoumalihoumed.github.io

OudoumAlihoumed oudoum-ali-houmed Google Scholar

Research Interests

AI safety and robustness of ML systems: adversarial robustness, reliability under distribution shift, and red-teaming and evaluation of frontier models and LLMs. Also interested in monitoring and control mechanisms, mechanistic interpretability, and neurosymbolic approaches to trustworthy AI.

Publications

[1] **O. A. Houmed**, O. Ceran. "A Systematic Review on Evolution, Challenges, and Future Trajectories of Endpoint Security: Integrating Zero Trust and Federated Learning Perspectives." *Artificial Intelligence Studies*, 2025. doi:10.30855/ais.2025.08.01.02

[2] **O. A. Houmed**, K. Jackson. "Can We Trust Deception Monitors for AI Agents? A Robustness-Gap Protocol and the Limits of Adaptive Evasion." *In preparation* – target: AI4Good Workshop @ NeurIPS. [code] [blog]

[3] **O. A. Houmed**. "Adversarial Threats to Safety-Critical Medical AI: A Security Assessment of 20 Deep Learning Tumour Detectors in Brain MRI and Kidney CT." *Manuscript*, ANÖROM Big Data & AI, 2025. [paper]

Education

M.Sc. Data, Knowledge and Hybrid Artificial Intelligence (DKAI) Sep 2025 – Present

Paris-Saclay University *Paris, FR*

- **Relevant coursework:** Towards Adaptive and Agentic AI; Exploring Data Declaratively; Towards Hybrid and Explainable AI; From Symbolic to Neurosymbolic AI

Bachelor of Science, Computer Engineering 2020 – 2024

Necmettin Erbakan University *Konya, TR*

- Graduated in the top 10% of cohort with a Full Scholarship

Additional Programs & Schools: ARENA (5-week intensive curriculum); AI Safety Fundamentals – Alignment Course & Technical AI Safety (BlueDot Impact); OxML 2025 (Oxford Machine Learning School)

AI Safety & Research Experience

Research Engineering Intern – Healthcare Generative Modeling & AI Robustness May 2026 – Sep 2026

Kappa Santé *Paris, FR*

- Spearheading applied research on constraint-guided virtual patient generation to address distribution shifts and augment underrepresented cohorts in longitudinal clinical datasets
- Designing conditional generation pipelines with explicit filtering for statistical fidelity and temporal coherence; evaluating downstream model safety under distribution shift (calibration, generalization, robustness)

AI Safety Research Fellow – Alignment & Security Track May 2026 – July 2026

Black in AI Safety & Ethics *Remote*

- Fellowship project under Krystal Jackson (UC Berkeley CLTC): a robustness-gap protocol measuring how much deception detection survives as adversary budget rises, evaluated on Llama-3.3-70B with the published Apollo residual-stream probe
- Built an adversary-budget ladder (b0–b4) with surface-classifier and chain-of-thought controls, plus a retention gate so a degraded agent is not mis-scored as successful evasion; findings intended to inform technical standards and policy

Research Assistant – AI Robustness Oct 2024 – Jun 2025

ANÖROM – Big Data & AI Dept. *Ankara, TR*

- Investigated the vulnerability of medical diagnostic architectures to adversarial perturbations using gradient-based optimization; implemented white-box attacks (FGSM, PGD) to quantify the robustness of tumour detection models
- Work developed into the sole-authored manuscript *"Adversarial Threats to Safety-Critical Medical AI"* [3]: 280 model–condition evaluations (20 architectures × 2 modalities), with Friedman/Nemenyi and Holm-corrected Wilcoxon validation

Additional Work Experience

Summer Research Intern Jul 2023 – Sep 2023

AI Security & Defence Lab (AISEC LAB) *Konya, TR*

- Co-architected Cyber Inspector, an ML pipeline covering data collection, constrained training (SVM, RF), and live deployment
- Co-authored *"Detection of Malicious Web Queries with Machine Learning"*: Random Forest over 1.3M labelled queries, 92.48% test accuracy, served live via Flask. [paper]

Selected Projects

SPEC-GAP – Indirect Prompt Injection in Multi-Agent Delegation May 2026 – Present

BASE Fellowship – presented at FAR AI | Llama 3.1 8B, LangGraph, PyTorch

- Built a LangGraph planner–worker–executor scaffold to study indirect prompt injection propagating through multi-agent delegation at 2- and 3-hop depths
- Trained linear probes on Llama 3.1 8B residual-stream activations to detect adversarial delegation, evaluating in- and out-of-distribution generalization with NARCBench-Core