Luis Ibáñez-Lissen PhD

28001 Madrid, España | LinkedIn | +34669768537 l.ibanezlissen@gmail.com

EDUCATION

University Carlos III - Madrid (Cum laude)

Phd studies. Computer sciences and technology. ***AI security and Safety*** *2022 - 2025*

1. *"On the Feasibility of Predicting Volumes of Fake News—The Spanish Case"*
   - *Journal:* IEEE Transactions on Computational Social Systems (Q1). Accepted
   - *Idea:* Novel approach on predicting the amount of fake news happening in a country using Different DNN models with an avg. error of 10%.
2. *"Use of transfer learning for affordable in-context fake review generation"*
   - *Journal:* IEEE Transactions on Big Data (Q1). Accepted.
   - *Idea:* Used different LLMs to generate in-context fake reviews that fools detectors and online users.
3. *"Matrix profile for privacy preserving continuous authentication"*
   - *Conference:* ARES Vienna Conference (Core B). Accepted
   - *Idea:* Novel approach in using incremental Matrix profile to perform continuous authentication of users in a privacy preserving way inline with sota accuracy (99%) .
4. *"Poisoning multitask multi-modal foundational generalistic models"*
   - Journal: Information system frontiers (Q1) Accepted . Analysis of multimodal models (LMMS) and transferability across tasks of backdoor attacks. Rebuttal period.
5. *"LPASS: Linear Probes as Stepping Stones for vulnerability detection using compressed LLMs*
   - Journal: JISAS. Accepted.
   - Idea: Novel Application of linear probes to find optimum interpretable ways to prune model layers while maintaining accuracy.
6. *"LUMIA: Unimodal and Multimodal membership inference Attacks."*
   - Conference: ESORICS (Core A). Accepted.
   - Idea: Novel Application of linear probes as a way to understand which samples were used during the training of LLMs and LMMS.
7. *"Large language models can learn and generalize steganographic chain-of-thought under process supervision."*
   - Accepted. NEURIPS (Core A*). Working with Cambridge university.
   - Idea: Using RL to elicit steganographic behaviours on LLMs while performing oversight on COT.

University Carlos III - Madrid

Double master on Computer informatics and CyberSecurity (Bilingual mode) 2020 - 2022

- Honors in Software for IoT and Seminars. Grade: 8.42/10 CS and 8.12/10 Cybersecurity.
- Master thesis of “AI for finding Vulnerabilities correlations” graded with 9.4/10.
- Used transformers, LLMs and NLP to find correlations between signatures and CWEs. Overall accuracy of the system 90%
- Master thesis of “Use of transfer learning for affordable in-context fake review generation” graded with a 10/10 and proposed for honors.

University Carlos III - Madrid

Computer Science and Engineering (Bilingual mode) 2016 - 2020

- Class representative, bachelor thesis proposed for honors. Grade: 7.59/10.

WORK EXPERIENCE

Banco Santander - Madrid

AI scientist December 2025 - Ongoing

- Working on the team in charge of transforming the internal processes of the Bank.
- Recently helping to create the group.
- Focus on AI security, RL and Governance.

University Carlos III - Madrid
Researcher (COSEC x INCIBE) *April 2022 - December 2025*

- Strategic national project on AI security and explainability with INCIBE.
- Research on XAI, NLP with LLMs, time-series forecasting, data poisoning techniques and time series classification.
- Finished DeproFake project, funded by the Community of Madrid. Focused on Misinformation and the impact of AI generated content over social media and AI systems.

FAR.AI- remote
Research collaborator *September 2025 - Ongoing*

- SPAR 25 mentee, on a Jailbreak project.
- Developing a toolkit to accelerate red teaming efforts on frontier models. Lead the design and implementation of the Multiturn and Long-Context archetypes.

INRIA - Paris
Invited Researcher November 2023 - January 2024, May 2024-July 2024, September 2025- Nov. 2025

- Invited researcher inside the PETRUS team at INRIA paris. Research on Privacy preserving continuous authentication.
- Research on Membership inference attacks for LLMs.
- Currently working on a Join paper with CISPA on Multiagents Fairness.

Advanet - Tokyo
Software developer *January 2023 - September 2023*

- Designed and implemented the software and firmware architecture for a Lora Micro-gateway.
- Implemented the firmware and the integration with the cloud of a Lora microcontroller.
- Designed an internal tool for Email summarization with LLMs. Deployed with podman and pytorch.

Bitcorp Creative Labs Italia - Remote
Software Developer *March 2020 - April 2022*

- Created and deployed a containerized LSTM autoencoder for anomaly detection on internal networks by analyzing key performance indicators on real time using Tensorflow, Docker, Flask and Python.

Sidertia Solution - Madrid
Machine learning intern *September 2021 - January 2022*

- Research about attention mechanisms and Transformers for signatures of software inventories with NLP and Transformer with an accuracy of 90%. Implemented in Tensorflow and Python.

EXTRA EXPERIENCE

Embassy of France in Spain Research Support - Paris

- Economical support to travel and work with Inria.

SPAR AI safety research - Remote

- Joint effort with FAR.AI to create a Toolkit for Jailbreak evals on Frontier models.

MARS AI safety research hub - (In place)

- Granted to Collaborate with Cambridge University to make research.

AI alignment course from Bluedot.org (remote)

- Granted to make an AI alignment course.

T3chfest: "Cybersecurity and AI: A love story" - Madrid
Invited talk April 2024

- Invited to talk about cybersecurity and AI at the T3chfest among 900+ proposals.

Vulcanus traineeship: "EU-JAPAN" - Tokyo

Grant *September-2022- September 2023*

- Conceded the “EU-JAPAN” Vulcanus grant internship starting in September 2022. Acceptance rate of $< 1\%$. Thanks to this, I attended the Naganuma School to learn Japanese in Shibuya to later work in a Japanese company.

Digitales: “Future voices” - Madrid

Invited talk July 2022

- Invited to talk at the summit with the Secretary of state in Science and innovation.

Google Tensorflow - Madrid

Certificate February 2021

- Successfully passed the Google official tensorflow developer certificate.

HarvardX TinyML - Madrid

Professional certificate *July 2021- August 2021*

- Successfully completed the HarvardX online course on TinyML.

Telefónica Sofía Challenge - Madrid

Hackathon November 2021

- Won the first National extended hackathon organized by the main Technological company in Spain by applying tiny machine learning on the edge of microcontrollers for remote sensing and sounds recognition.

SKILLS & LANGUAGES

Skills: Python, Java, C, Tensorflow, Pytorch, Machine Learning, SQL, MongoDB, Docker, Linux, Metasploit, Git, IoT, Spark, Hadoop, NLP, Transformers, LLM, LMM, XAI, AI safety.

Languages: Spanish (Native), English (Professional), Japanese (B1).

Ongoing research

8. *"Gender Bias and Influence on SA and MAS scenarios"*
    - *Idea:* Understanding gender bias on MAS scenarions.
    - Collaborative research with CISPA and INRIA.
9. *"Detection of Lies and Deception in LLMs using Residual Stream Activations with scalable oversight"*
    - *Idea:* Understanding generalization on Linear probes. Testing an scalable oversight approach to generate synthetic data. We want to understand how synthetic data can generalize across datasets to catch deceptive behaviours.