



Fran Cortés

LinkedIn | francordel.com | UMU profile | GitHub |
franciscojose.cortesd@um.es

Research Interests

LLM red-teaming and adversarial robustness; safety and security of LLM agents and multi-agent systems; jailbreaks, prompt injection, and the design of mitigations against hard-to-detect failure modes.

Publications & Research

- **Cortés-Delgado, F. J., Rodríguez, et al.** (2026). ContextualJailbreak: Evolutionary Red-Teaming via Simulated Conversational Priming. arXiv:2605.02647. An evolutionary black-box jailbreak over simulated multi-turn dialogue, with an in-loop two-stage judge. Across 50 representative HarmBench behaviors it reaches 100% ASR on gpt-oss:20B and 90% on gpt-oss:120B; the 40 maximally harmful attacks found transfer without adaptation to closed frontier models (90% gpt-4o-mini, 70% gpt-5, 70% gemini-3-flash). Co-authored with researchers from IBM Research Dublin and the University of Murcia.
- Implemented the above as an open-source red-teaming strategy plugin in IBM ARES (AI Robustness Evaluation System), a framework for orchestrating automated red-teaming of LLMs and agentic AI systems. github.com/IBM/ares/pull/521
- During my research visit at Cambridge (CaMLSys), contributed an open-source implementation of FedRDF (a robust, FFT-based aggregation method defending federated learning against poisoning attacks, originally proposed by E. Mármol Campos et al., UMU) to the Flower framework; PR under review. github.com/adap/flower/pull/6044
- **Cortés-Delgado, F. J., Mármol Campos, E., Hernández-Ramos, J. L., Skarmeta, A., et al.** (2026). Enhancing security and privacy in federated learning for distributed systems: The REMINDER approach. In Proc. Global IoT and Edge Computing Summit (GIECS 2025), pp. 57–70. Springer. Outcome of the EU CHIST-ERA REMINDER project. doi.org/10.1007/978-3-032-09555-8_4
- Ongoing (with IBM Research Dublin and Cambridge / CaMLSys): agentic multi-agent attack, in preparation. Current results reach 100% ASR on gpt-oss:20B and 98% on gpt-oss:120B, with transfer to closed frontier models (ASR@4 of 58.8% on gpt-5, 52.9% on gpt-5.4-nano, 41.2% on gpt-5.5, 32.4% on claude-opus-4-8).
- **Cortés-Delgado, F. J.** (2026). University of Murcia. Master's thesis, graded 9.9/10 (Sobresaliente). Designed and implemented an agentic framework for the automated red-teaming of LLMs, orchestrating multiple cooperating agents to discover and evaluate jailbreak strategies at scale, extending the evaluation methodology used in the ContextualJailbreak line of work.

Research Experience

Visiting Researcher — University of Cambridge (CaMLSys Lab)

Cambridge, UK · 09/2025

- Research on agentic red-teaming of LLMs, identifying vulnerabilities and evaluating model robustness, with Prof. Nic Lane's group; collaboration ongoing.

Visiting Research Engineer — IBM Research

Dublin, Ireland · 05/2025–08/2025

- Opened an LLM red-teaming research line under Dr. Stefano Braghin; designed an agentic red-teaming framework and contributed the ContextualJailbreak strategy to IBM ARES; collaboration ongoing.

AI & Cybersecurity Researcher — University of Murcia

Murcia, Spain · 10/2024–present

- Work on the EU CHIST-ERA REMINDER project (privacy-preserving federated learning, robust to poisoning and inference attacks), a four-country consortium with Siemens, AIT Austrian Institute of Technology (Vienna), and UWE Bristol; also adversarial AI and LLM-based vulnerability identification, under Prof. Antonio Skarmeta and Prof. José Luis Hernández-Ramos.

Industry Experience

Specialist Software Engineer — Vodafone

Málaga, Spain · 09/2023–10/2024

- Built production microservices (Java Spring Boot, MongoDB, Kafka, Docker) in an agile environment; led graduate technical hiring and mentored interns.

Deep Learning Engineer Lead — Bemyvega

Murcia, Spain · 12/2022–05/2023

- Led the AI team building real-time transcription systems with NLP and deep learning (incl. LLMs); model selection per use case.

Co-Founder & CEO — MiSintaxis

Murcia, Spain · 09/2019–09/2023

- Built the world's #1 syntax-teaching app (85,000+ downloads). Fine-tuned BLOOM and GPT-2 on the AnCora corpus for Spanish constituency parsing (strong F1); deployed on AWS with Hugging Face/PyTorch. Led a team of 4 and the full business side.

Education

MSc in Artificial Intelligence — University of Murcia

Murcia, Spain · 09/2025–07/2026

- Overall GPA: 9.14/10. Master's thesis — an agentic framework for the automated red-teaming of LLMs — graded 9.9/10 (Sobresaliente).

BSc Computer Engineering, with Honors — University of Murcia

Murcia, Spain · 09/2019–07/2023

- GPA 8.6/10, with Honors (top 1% of the university). Final-year research thesis on NLP with LLMs graded 10/10. Erasmus semester at Politechnika Wroclawska, Poland (2021–2022), taught in English.

Awards

- Extraordinary Prize for Academic Excellence in Computer Engineering (University of Murcia) — best graduate of the cohort, EURO-INF accredited programme.
- Best Final Degree Project Award (€1,500) — NLP + deep learning for Spanish grammar; led to the creation of MiSintaxis.
- Extraordinary High School Award (Region of Murcia) — top student, all subjects passed with honors in the Science track; award covered the full first year of university tuition.
- First place in multiple startup/entrepreneurship competitions with MiSintaxis.

Skills

- Python · PyTorch · Hugging Face · NLP · Deep Learning · LLM red-teaming · adversarial ML
- Java Spring Boot · Docker · AWS · Git · MongoDB · MySQL · C/C++ · JavaScript · SCRUM · BDD
- Languages: Spanish (native); English — C1 Advanced (Cambridge CAE, Grade B, 2025); French — DELF B2