Houssem Guermazi

MVA x **CentraleSupélec**

+33 6 78 81 16 54 • houssem.guermazi@student-cs.fr • houssem-guermazi • houssem.site

Profile

Research-driven machine learning engineer pursuing the MVA Master at ENS Paris-Saclay and the CentraleSupélec Engineering degree, with a strong mathematical and experimental background I am particularly interested in training, and improving modern machine learning models through rigorous experimentation. I am seeking an end-of-study internship with a strong research component, starting in April 2027, where I can contribute to technically ambitious machine learning problems and work closely with research and engineering teams.

Current Research Project

Token-Level Sensitivity in Lossy Speculative Decoding (MICS) Oct 2026 – Present

- Investigating where approximate token acceptance is safe in LLM speculative decoding, characterizing the quality–speed trade-off of lossy verification rules against the exact, distribution-preserving baseline.

Experience

R&D / AI Intern – *BNP Paribas Innovation Lab* Jan 2026 – Jul 2026

- Designed and implemented an end-to-end multi-agent system from scratch using LangChain and LangGraph to automate quarterly financial press-release generation from 300–600 Word documents, with conditional verification loops preserving document-level consistency and an estimated 1–2 week reduction in time-to-first-draft.
- Built a hybrid RAG pipeline combining BM25 and dense semantic retrieval over ChromaDB to ground each generated section in financial evidence collected from multiple BNP Paribas entities.
- Designed a claim-level factuality evaluation framework in which multiple LLM judges assessed whether generated statements were explicitly supported in the source documents, using an 80% consensus threshold.
- Built an OpenAI-compatible inference gateway across heterogeneous Azure-hosted LLMs, implementing model-specific JSON schema adapters, tool-call serialization, reasoning configurations, model routing, load balancing, and capacity controls to reliably support agentic and coding workloads.

Data Scientist – *SFR* Jul 2025 – Dec 2025

- Improved a large-scale fraud detection pipeline using XGBoost and Random Forest, addressing severe class imbalance through stratified sampling, class weighting, and threshold optimization.
- Developed hybrid pipelines combining embeddings extracted from deep neural networks and Vision Transformers with XGBoost and Random Forest classifiers, leveraging learned representations alongside robust tabular models.
- Modeled relational interactions as graphs and investigated graph-based learning, neighborhood sampling, scalability, and generalization to previously unseen entities and fraud patterns.
- Developed a multi-agent system on Vertex AI, integrating specialized NLP and retrieval agents through A2A/MCP-inspired communication protocols and RAG over enterprise documentation.

Research Projects

GPU-Accelerated 3D Delaunay Tetrahedralization (MICS x Safran) Sept 2026

- Conducted a large-scale empirical study of seven CPU/GPU 3D Delaunay algorithms on point clouds of up to one million points using an NVIDIA A100, evaluating computational performance, scalability, and numerical robustness.
- Developed a rigorous validation framework based on robust geometric predicates, convex-hull coverage, volume consistency, and tetrahedral Jaccard similarity; introduced controlled jitter to reduce exact-arithmetic bottlenecks and achieved speedups of up to 16.7× for gDel3D and 22.4× for gStar4D.

Research Hackathon on Trustworthy AI (CISPA) Nov 2025

- Placed 3rd among 16 teams in the CISPA Paris Cybersecurity & AI Hackathon and qualified for the European Grand Final.
- Developed minimal-L_2 black-box adversarial attacks using surrogate CNN/ViT ensembles, iterative PGD variants, transfer attacks, and binary-search perturbation calibration against a hidden image classifier.
- Built generative-model attribution features from multi-autoencoder reconstruction errors across VAR, RAR, Stable Diffusion, and Taming Transformers, and fine-tuned a ResNet-50 watermark detector with ROC-based threshold calibration for TPR at 1% FPR.

Deep Reinforcement Learning for Nonlinear Control (NUS) Jan 2025 – May 2025

- Implemented a Deep Q-Network agent to balance an inverted pendulum in a simulated CartPole environment.
- Applied experience replay, ε-greedy exploration and temporal-difference updates to learn a control policy.

Cross-Attentive Multimodal Time-Series Forecasting (NUS) Jan 2025 – May 2025

- Designed an attention-enhanced Seq2Seq architecture for multi-horizon box-office forecasting, using an LSTM encoder to fuse historical revenue sequences, static numerical features, and learned categorical embeddings, with an LSTM decoder conditioned on positional and temporal embeddings.
- Integrated multi-head cross-attention between decoder queries and encoder states, together with residual connections, layer normalization, dropout, and a feedforward prediction head; incorporated ACF/PACF-selected lags, cyclical time encodings, and rolling-window features, achieving a 34.51% test MdAPE.

Transformer-Based Deep Metric Learning for Fashion Retrieval (CS x QuickFind) Sep 2024 – Dec 2024

- Designed and fine-tuned a weight-shared deep metric-learning architecture using ViT and SwinV2 encoders to map anchor, positive, and negative fashion images into a discriminative latent space optimized with triplet-margin loss.
- Compared ViT global patch self-attention with SwinV2 hierarchical shifted-window attention, extracting last-hidden-state embeddings for cosine-similarity retrieval and benchmarking representation quality against Jina CLIP through distance distributions, t-SNE projections, and client-defined visual attributes.

Movie Recommendation Website (CS x Theodo) May 2024

- Implemented a recommendation system and web platform using React, Node.js, and REST APIs.
- Studied latent representations, sparse matrices, and optimization methods underlying recommendation quality.

Promptable Segmentation and Video Tracking (CS x Safran) Jan 2024 – Jun 2024

- Fine-tuned and benchmarked YOLOv5 and YOLOv8 on KITTI and crack-segmentation datasets, evaluating precision, recall, F1, and mask quality.
- Combined YOLOv8 instance segmentation with SAM's image, prompt, and mask encoders for point-guided annotation and persistent-ID video tracking.

Suicidal Risk Detection from Tweets (CS) Oct 2023

- Built a classification pipeline achieving 90% accuracy using Random Forest, Word2Vec, and TF-IDF.
- Performed NLP preprocessing, feature engineering, and model evaluation.

Education

Master Mathematics, Vision and Learning (MVA), *M2, ENS Paris-Saclay* 2026 – 2027
Research-intensive training in modern machine learning , with emphasis on large-scale optimization and foundation models.

CentraleSupélec, *M2, Engineering Degree* 2023 – 2027
First-year GPA: 4.23/4.33
Second-year GPA: 4.03/4.33
Specialization in Artificial Intelligence, applied mathematics, and research-oriented machine learning.

National University of Singapore (NUS), *Exchange Semester* Jan 2025 – May 2025
GPA: 4.0/4.0
Advanced coursework in Artificial Intelligence, including Advanced Deep Learning, Reinforcement Learning, and Fundamentals of Machine Learning.

IPEST, *Intensive Preparatory Classes (MPSI / PSI*)*, Tunisia 2021 – 2023
Highly selective scientific preparatory track with intensive training in mathematics and physics.

Technical Skills

- **Machine Learning :** PyTorch, scikit-learn, Transformers, ViT, Swin Transformer, ResNet, CNNs,
- **LLM & Retrieval Systems:** LLM inference, speculative decoding, Hugging Face Transformers, RAG, multi-agent systems, LangChain, LangGraph, Chroma
- **GPU & Scientific Computing:** Python, NumPy, SciPy, SLURM, parallel computation, scalability analysis
- **MLOps & Cloud:** GCP Vertex AI, Azure, Docker, model serving, model routing, load balancing
- **Graph ML:** graph-based machine learning, Neo4j, Cypher, NLP, multimodal modeling, fraud detection
- **Software Engineering:** Git, FastAPI, REST APIs, React, Node.js

Languages

French (C2), **English** (C1), **Arabic** (fluent), **German** (B2)