Ilyass Ouardi

M.Sc. Student in Computer Science · Aspiring Ph.D. Researcher (AI Safety & Mechanistic Interpretability)
Milan, Italy · (+39) 320 175 5208 · ilyassouardi@gmail.com · ouardi-ilyass.github.io · GitHub · LinkedIn

Research Interests

My research focuses on **mechanistic interpretability**, **representation geometry**, and **verifiable AI safety**. Rather than treating interpretability as post-hoc inspection, I approach it as a foundational empirical science to make internal model mechanisms verifiable, steerable, and robust on the path toward safe AI. My current work investigates how abstract concepts—such as propositional truth, refusal, and ethical norms—are geometrically structured within LLM activation manifolds: (1) *Model Biology*: whether concepts conform strictly to 1D linear directions or organize into higher-dimensional non-linear concept cones and multi-dimensional submanifolds; (2) *Surgical Representation Interventions*: whether representation editing (INLP, LEACE, activation addition) is truly surgical, designing diagnostic protocols to measure off-target side-effects as rigorously as target changes; and (3) *Fairness & Linguistic Representation*: how diverse and low-resource linguistic concepts are encoded and preserved across activation spaces.

Education

M.Sc. in Computer Science, University of Milan, Italy — Sep 2024 – Present
Thesis: *The Geometry of Absence, Opposite, and the Activation Manifold in Representation-Level Interventions on LLMs*
Advisors: Prof. Alfio Ferrara & Elisabetta Rocchetti (PhD) · ISLab (Information Systems & Knowledge Management Lab)
Weighted average: **29.13/30** (3.88/4.00) · Expected completion: Summer 2027
Selected Coursework: Natural Language Processing (30 *cum laude*), Distributed and Pervasive Systems (30 *cum laude*), Security of Service-Oriented Architectures (30 *cum laude*), Advanced Biometrics (30), Audio Pattern Recognition (30), Privacy and Data Protection (29).

B.Sc. in Computer Science, University of Milan, Italy — Sep 2020 – Oct 2024
Thesis: *Design and development of a donation-tracking system for RECUP APS*
Advisor: Prof. Carlo Maria Bellettini
Final Grade: **110/110 *cum laude*** · Weighted average: **28.55/30** (3.81/4.00)
Mathematical & Theoretical CS Foundations: Discrete Mathematics (30 *cum laude*), Mathematical Logic (30 *cum laude*), Automata and Formal Languages (30 *cum laude*), Programming Languages (30 *cum laude*), Artificial Intelligence I (30 *cum laude*), Calculus (30), Statistics and Data Analysis (30), Algorithms and Data Structures (30).

Diploma in Computer Science (Perito Informatico), ITT "G. e M. Montani", Fermo, Italy — Sep 2015 – Jul 2020
Final Grade: **100/100 *cum laude***
Specialization in Software Engineering, Algorithms, and Computer Networks.

Research Experience

Independent Research — The Multidimensional Geometry of Truth in LLMs — Dec 2025 – Jul 2026
ISLab, University of Milan · Sole author (NLP Research Track)
- Investigated whether propositional truth in LLMs is mediated by a single linear vector or by a higher-dimensional geometric cone. Adapted Wollschläger et al. (2025) to formulate **Truth Direction Optimization (TDO)** and **Truth Cone Optimization (TCO)**: gradient-optimized bases constrained under causal necessity, sufficiency, and KL-retention losses.
- **Findings:** Proved that a 1D linear probe fails to reliably capture truth in larger models, whereas a low-dimensional cone ($k = 2$) captures truth representations substantially better. Discovered that the cone's basis is near-orthogonal to the Difference-in-Means probe, yet individually causally steers factual generation.
- **Manuscript:** *The Multidimensional Geometry of Truth in LLMs: A Concept-Cone Extension of the Linear Representation Hypothesis* (2026). [Draft] [Code]. PyTorch & `nnsight` on single A40 GPU.

Research Project — What Does a Heart-Sound Classifier Actually Learn? — Apr 2026 – Jul 2026
University of Milan · Sole author (Audio Pattern Recognition Track)
- Evaluated three acoustic representations (MFCC, log-Mel, wavelet packets) across three model families (SVM, Random Forest, CNN) on the PhysioNet/CinC 2016 benchmark, pairing every performance claim with an adversarial refutation protocol.
- **Findings:** Demonstrated that a headline MAcc of 0.866 is largely an artifact of recording-site confounding: site was recoverable from representations at 0.647 (vs. 0.167 chance), and diagnostic accuracy collapsed from 0.840 to 0.457 under leave-one-site-out validation. Proved that rigorous diagnostic audits are essential before attributing semantic capability to neural models.
- **Manuscript:** *What Does a Heart Sound Classifier Actually Learn?* (2026). [Draft] [Code].

Research Project — The Legend Challenge: Ethical Compliance in LLMs — Feb 2026 – Jun 2026
SESAR Lab, University of Milan · With F. Bylyshi & H. El-Khazri (Equal contribution)
- Tested whether language models internalize behavioral norms more effectively from explicit regulatory rules or from narrative stories depicting compliance in action.
- Engineered a controlled corpus generation pipeline and constructed **GenderEqGLUE**, a five-task evaluation benchmark on held-out documents. Proved that rule-based and narrative-based training instill complementary capabilities.
- **Manuscript:** *The Legend Challenge: Embedding Ethical Compliance into LLMs through Champion Narratives* (2026). [Draft] [Code].

Teaching & Academic Service

Teaching Assistant, *Accertamento delle Competenze Informatiche* (3 CFU), University of Milan Jan 2026 – Present
Scientific Lead: Prof. Stefano Montanelli · Delivered across university degree programs by CTU

- Prepare and deliver technical lectures on computing fundamentals, databases, and algorithmic thinking to undergraduate classes; author accompanying pedagogical materials and lab exercises.
- Supervise hands-on programming laboratory sessions and invigilate institutional examinations.

Selected Engineering Experience

Lead Developer / Consultant, RECUP APS — University of Milan Jan 2024 – Dec 2025

- Designed and deployed an end-to-end cloud platform for food-waste redistribution (React, TypeScript, Supabase PostgreSQL, Netlify Functions) serving markets across Milan and Rome; implemented multi-role access control (RBAC).
- **RecupGPT** (2025): Developed an intelligent natural-language interface over the production database using LangChain, LangGraph, and NL2SQL with automated query rewriting and semantic schema routing.

Full-Stack Developer Consultant, Value Process Srl, Milan Jun 2022 – Sep 2022

- Developed microservices architecture in Spring Boot and React; participated in migrating legacy systems to serverless.

Fellowships, Honours & Leadership

Full Academic & Merit Scholarships (Tuition and housing coverage), University of Milan Sep 2020 – Sep 2026
Lead the Future — Mentee in premier STEM mentorship community for high-achieving Italian scholars Sep 2025 – Present
BOOST '24 — Selected for Bologna Orthogonal Summer Term for top B.Sc./M.Sc. Computer Science students Aug 2024
Futuro Annunciato '24 — Selected by University of Milan for competitive student cohort Jul 2024
Nomination, "Alfieri del Lavoro" Award — Nominated by Presidency of the Italian Republic (Top 25 high school grads) Jul 2020
National Olympiads in Informatics & Mathematics — Qualified school finalist and competitor 2017 – 2020

Technical Skills & Languages

Mechanistic Interpretability & Safety: Denoising activation patching, directional ablation & addition, Difference-in-Means probes, concept cones (TDO/TCO), concept erasure (INLP, LEACE), activation steering, causal mediation analysis, SHAP, Grad-CAM.
Experimental Methodology: Causal mediation protocols, permutation/analytic null testing, grouped/leave-one-group-out cross-validation, multiple-comparison correction, bootstrap estimation, adversarial null hypothesis testing.
ML Tooling & Frameworks: PyTorch, `nnsight`, Hugging Face (Transformers, Datasets, Accelerate), scikit-learn, LangChain, LangGraph.
Systems & Languages: Python (proficient), Java (proficient), TypeScript/JavaScript (proficient), SQL/PostgreSQL; C, Go, Scala.
Languages: Italian (native), Arabic (native / bilingual), English (fluent, full professional proficiency).

References available upon request · Full code repositories and research manuscripts available at `ouardi-ilyass.github.io`