Abir HARRASSE

abirharrasse.github.io | +212 688 972327 | Abir.HARRASSE@emines.um6p.ma

Education

Master's Degree in Industrial Management — *EMINES School of Industrial Management*
Minor in Data Science: ***GPA: 3.89/4.0*** — *September 2024 - September 2025*
Coursework includes: NLP, LLMs, Machine Learning, Optimization, Statistics.

Bachelor's Degree in Industrial Management — *EMINES School of Industrial Management*
Engineering degree program: ***GPA: 3.96/4.0*** — *September 2020 - September 2024*
Coursework includes: Real Analysis, General Topology, General Algebra, Linear Algebra, Multivariable Calculus, Martingales, Lebesgue Integration, Measure Theory, Probabilities, Linear Programming.

Papers

Tracing Multilingual Representations in LLMs with Cross-Layer Transcoders
Abir Harrasse, Florent Draye, Zhijing Jin, Bernhard Schölkopf
(Under Review) Available at: temporary drive link

TinySQL: A Progressive Text-to-SQL Dataset for Mechanistic Interpretability Research
Abir Harrasse, Philip Quirke, Clement Neo, Dhruv Nathawani, Amir Abdullah
(EMNLP Main Conference) Available at: arXiv:2503.12730

Disentangling and Steering Multilingual Representations: Layer-Wise Analysis and Cross-Lingual Control in Language Models
Abir Harrasse, Florent Draye, Bernhard Schölkopf, Zhijing Jin
(ICML AIW Workshop) Available at: AIW-Workshop

Activation Space Interventions Can Be Transferred Between Large Language Models
Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, **Abir Harrasse**, Amirali Abdullah
(ICML 2025) Available at: arXiv:2503.04429

Industry Experience

Martian Learning Inc. — *November 2025 - Present (Incoming)*
Research Fellow — *Bay Area, San Francisco*

- Will continue mechanistic interpretability research and develop uncertainty estimation methods for model routing.

Martian Learning Inc. — *September 2024 - March 2025*
Research Intern (Remote) — *Bay Area, San Francisco*

- Received return offer for full-time Research Fellow position.
- Developed mechanistic interpretability framework to probe models' causal reasoning algorithms and transfer safety interventions across models. Work accepted at ICML 2025 and EMNLP 2025.

Research Experience

Causal Interpretability - The Empirical Inference Lab — *April 2025 - September 2025*
Research Intern, supervised by ***Prof. Zhijing Jin*** — *Max Planck Institute for Intelligent Systems*

- Focused on studying and interpreting multilingual mechanisms in Large Language Models.
- Developed new tools for training data attribution estimation.

LLMs evaluating LLMs — *April 2024 - August 2024*
Research Intern, supervised by ***Prof. Chaithanya Bandi*** — *National University of Singapore*

- Designed adversarial multi-agent systems (roles: advocates, judges) for LLM output evaluation.

- Improved decision-making, achieving an 8% performance boost over prior methods.

RL with generative models *January 2023 - January 2024*
Research assistant, supervised by ***Prof. Omar Saadi*** *College of Computing-UM6P, Benguerir*

- Focused on model-based reinforcement learning using generative models.
- Tested existing algorithms and optimized their runtime and sampling efficiency.

Talks

Training Data Attribution *August 2025*
Amazon *Tübingen, Germany*

- Presented training data attribution in LLMs, covering verbatim and gradient-based methods, their strengths, limitations, and open challenges..

Awards

First Class Honours, EMINES, UM6P	Graduated with first-class honours for exceptional academic performance.
Best Master's Thesis, EMINES, UM6P	Recognized for outstanding research on understanding and interpreting language processing in LLMs.
National Moroccan Merit Scholarship FAR, 2020	Awarded to the top 50 national scores in high school final examinations, representing the top 0.1% of all candidates.
UM6P Excellence Scholarship, 2020-2025	Granted for exceptional performance in entrance examinations, placing within the top 1% of applicants.

Key Achievements and Projects

BCG Platinion Hackathon *October 2023*
National Winner, Internationl 2nd Place *BCG Casablanca*
Our mobility solution placed 2nd internationally at the hackathon (150+ participants), optimizing transport matchmaking using Google's VRP-Solver.

NaMO - Preparing for IMO's 61st edition *January 2019 - March 2020*
Participant *Rabat, Morocco*
Selected among the top 25 in the country's most prestigious math competition, gathering the best high school students to prepare for international contests like the IMO.

Skills

Languages	*Native:* Arabic, French. *Fluent:* English. *Beginner:* Mandarin
Programming Languages/Tools	Python, Pytorch, TransformerLens, SAELens, Nnsight, Transformers, Diffusers, Datasets, Slurm, HTCondor

Volunteering

MathMaroc *September 2021 – September 2024*
Vice General Secretary
Organized events, workshops, and outreach programs to engage students in mathematics.

The GenAI Winter School *October 2023 – February 2024*
Organizer
Coordinated the **GenAI Winter School**, featuring distinguished researchers such as **Yann LeCun** and **Eric Xing**.