Wassim (Wes) Bouaziz

wassim.bouaziz@mistra.ai Google Scholar wesbz wassim-bouaziz

Research Experience

Mistral AI *AI Scientist* **Nov. 2025 – Present**
Mistral AI, Safety team *Paris, France*
Initiating research projects on AI security.

Meta & École polytechnique, X *Research Assistant, PhD student* **May 2022 – Oct. 2025**
FAIR & CMAP department at École polytechnique *Paris, France*
I explored a spectrum of research questions at the intersection of AI Security, Safety, and Interpretability.
- Published research work (either as first author or contributor) in top venues including ICLR, ICASSP and NeurIPS.
- Contributed to Llama 3 models by developping safety evaluations and safety data annotation campaigns.

Independant *Machine Learning Consultant* **Jan. 2021 – May. 2022**
Consulting and Teaching in Machine Learning *Paris, France*
- Giving Deep Learning lectures and supervising projects at Le Wagon.
- Advising startups on project feasibility and state of the art in machine learning research on various topics.

Facebook AI Research *Research Intern* **Jul. – Nov. 2020**
Noise-robust Automatic Speech Recognition *Paris, France*
- Achieved 10% accuracy improvement using audio data augmentation to enhance robustness of speech recognition systems.

ENS *Research Intern* **Apr. – Aug. 2019**
ENS CoML team *Paris, France & Montréal, Canada*
- Classification of child/adult directed speech for a world-wide study on child language acquisition (ACLEW).

CentraleSupélec *Graduate Research Assistant* **Oct. 2018 – Apr. 2019**
Graphs, Algorithms and Combinatorics team *Saclay, France*
- Contributed to improving the known time complexity of minimum cuts counting in graphs.

Snips (Acquired by Sonos) *Embedded Software Intern - AI* **Jun. – Aug. 2018**
Embedded systems - Real time inference *Paris, France*
- Developed real time inference application using Snips' voice assistant with a $< 30\%$ CPU usage on a Cortex-M7 microcontroller.

Education

Meta & École polytechnique, X *PhD in Mathematics and Computer Science* **May 2022 – Oct. 2025**
Defended Dec. 2025. Towards Secure and Trustworthy Machine Learning:
From Data Poisoning to Ownership Verification *Paris, France*
- Advisor: Eric Moulines & El Mahdi El Mhamdi (CMAP, École polytechnique), Nicolas Usunier (Helsing), Max Nickel (FAIR, Meta)

Paris-Saclay University *MSc in Mathematics, Vision and Learning ("MVA")* **2019 – 2020**
Among the World's best university in Mathematics in Shangai's ranking - with Highest Honors, GPA: 4.0/4.0 *Paris, France*

Université Paris-Saclay *MSc degree in Computer Science and Machine Learning* **2018 – 2019**
Double degree with CentraleSupélec, GPA: 4.0/4.0 *Paris, France*

CentraleSupélec *MEng degree in Mathematics and Computer Science.* **2016 – 2019**
Top Engineering school in Computer Science and Electrical Engineering. GPA: 3.9/4.0 *Gif-sur-Yvette, France*

Preparatory classes *Lycée Blaise Pascal* **2014 – 2016**
Two-year undergraduate intensive program for the entrance exams to top French Engineering schools *Orsay, France*

Technical Skills

AI Large scale training, Language models training (pre & post training),
Programming Python (PyTorch), Bash, Latex, C++$^{(-)}$ SQL$^{(-)}$, HTML/CSS/JS$^{(-)}$,
Languages French (*native*), English (*C1*), Spanish (*B2*), Arabic (*A2*), Japanese (*A1*)

External Engagement

Reviewer ICLR (2023, 2025, 2026), NeurIPS (2024, 2025)

Invited Talks

- *"Risks of AI agents in high-stakes environment"* in the "House of Finance Days" seminar at Université Dauphine-PSL (2026).
- *"Towards Secure and Trustworthy Machine Learning: From Data Poisoning to Ownership Verification"* in the "AI for social good" course of the "AI and society" PSL Master (2026).
- *"AI security & data poisoning"* for *Hacienda*, a French tech community (2026).
- *"How can AI models be hacked?"* in the French Underscore_ tech show (2025) – >400k views on YouTube
- *"Harnessing AI's Fragility: A Tool for Dataset Tracing"* in PurrLab seminar (2025), INRIA Sierra seminar (2025), Callyope research seminar (2025)
- *"AI Security & Safety"* in the "Reliable and Collaborative learning" course at École polytechnique (2024)
- *"Digital commons in AI research"* in the "Introduction to digital commons and public policy making" course for Public Policy Master students at Sciences Po Paris (2023, 2024)

Publications

– First author publications

Byzantine Machine Learning: MultiKrum and an optimal notion of robustness — **October 2025**
*Gilles Bareilles, **Wassim (Wes) Bouaziz**, Julien Fageot, El Mahdi El Mhamdi* — *arXiv*

Winter Soldier: Hypnotizing Language Models at Pre-Training with Indirect Data Poisoning — **April 2025**
***Wassim (Wes) Bouaziz**, Mathurin Videau, Nicolas Usunier, El Mahdi El Mhamdi* — *ICLR 2026*

Data Taggants: Dataset Ownership Verification via Harmless Targeted Data Poisoning — **October 2024**
***Wassim (Wes) Bouaziz**, Nicolas Usunier, El Mahdi El Mhamdi* — *ICLR 2025*

Targeted Data Poisoning for Black-Box Audio Datasets Ownership Verification — **October 2024**
***Wassim (Wes) Bouaziz**, El Mahdi El Mhamdi, Nicolas Usunier* — *ICASSP 2025*

Inverting Gradient Attacks Makes Powerful Data Poisoning — **October 2024**
***Wassim (Wes) Bouaziz**, El Mahdi El Mhamdi, Nicolas Usunier* — *TMLR Dec. 2025*

Clustering Head: A Visual Case Study of the Training Dynamics in Transformers — **December 2024**
*Ambroise Odonnat, **Wassim (Wes) Bouaziz**, Vivien Cabannes* — *arXiv*

Easing Optimization Paths: a Circuit Perspective — **December 2024**
*Ambroise Odonnat, **Wassim (Wes) Bouaziz**, Vivien Cabannes* — ***Lecture** ICASSP 2025*

– Collaboration publications

On Monotonicity in AI Alignment — **May 2025**
*Gilles Bareilles, **Wassim (Wes) Bouaziz**, Sadegh Farhadkhani, Julien Fageot, Lê-Nguyên Hoang, Sébastien Rouault, Peva Blanchard, El-Mahdi El-Mhamdi* — *Under review*

On the strength of goodhart's law — **May 2025**
*Adrien Majka, **Wassim (Wes) Bouaziz**, El Mahdi El Mhamdi* — *2nd MoFA Workshop at ICML 2025*

Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset — **May 2025**
*Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, **Wassim (Wes) Bouaziz**, Jack Kussman, Manon Revel, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, Maximilian Nickel* — ***Oral** 2nd MoFA Workshop at ICML 2025*

The Llama 3 Herd of Models — **July 2024**
Llama Team, AI @ Meta — *arXiv*

Iteration head: A mechanistic study of chain-of-thought — **December 2024**
*Vivien Cabannes, Charles Arnal, **Wassim Bouaziz**, Xingyu Yang, Francois Charton, Julia Kempe* — *NeurIPS 2024*

On the Parameterized Complexity of Counting Small-Sized Minimum (S, T)-Cuts — **December 2023**
*Pierre Bergé, **Wassim Bouaziz**, Arpad Rimmel, Joanna Tomasik* — *SIAM Journal on Disc. Math.*