Clément DUMAS

Paris, France

clement[dot]dumas[at]ens-paris-saclay[dot]fr

linkedin/Clément Dumas | Google scholar | github.com/Butanium | Website: butanium.github.io

Neel Nanda's MATS 7.0 scholar with experience in mechanistic interpretability and model diffing. Publications at ACL 2025 (main conference) and workshops at ICML 2024 and ICLR 2025, interested in AI safety research positions.

Education

- 2024-2025: Computer Science research Master MVA (Mathematics, Vision, Learning) at **ENS Paris-Saclay**. The ENS is a selective institution that trains teachers and researchers
- 2023-2024: Computer Science research Master MPRI at **ENS Paris-Saclay**.
- 2022-2023: Double Bachelor's degree in Computer Science at **ENS Paris-Saclay**
- 2020-2022: Completed "classes préparatoires", an intensive two-year programme in the sciences with 12 hours of math per week, preparing for the competitive entrance exams to the ENS

Research Experience

✨: led to a first-author publication

- Since May 2025: MATS extension in Neel Nanda's stream. I was awarded a 1-year scholarship to continue my model diffing research
- Winter 2025 ✨: MATS 7.0 scholar in Neel Nanda's stream working on model diffing. An early version of our paper Overcoming Sparsity Artifacts in Crosscoders to Interpret Chat-Tuning was published in the ICLR SLLMs workshop
- October 2024: Completed the Neel Nanda's MATS stream training phase. It ended with a 2 weeks research sprint where we replicated and extended the Crosscoder paper. See our demo Colab
- Summer 2024 ✨: 5-month research internship at EPFL with Robert West and Chris Wendler. Our work, Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers , was spotlight at **ICML 2024 mechanistic interpretability workshop** and accepted at **ACL 2025** main conference
- July 2024: Attended the Human-aligned AI Summer School
- January 2024: Explored the emergence of XOR features in LLMs and the RAX hypothesis developed by Sam Marks. See our code
- October 2023 - May 2024: Supervised Program for Alignment Research (SPAR) under the supervision of Walter Laurito. We tried to apply Contrast-Consistent Search to Reinforcement Learning models .
- Summer 2023: Two months research internship with Jobst Heitzig on Aspiration-Based Q-Learning
- 2022-2023: Participated in “Séminaire Turing”, an AI alignment reading group at ENS Paris-Saclay
- December 2022: Participated in the AI testing hackathon organized by Apart Research. Our submission about Trojans in transformers was ranked #4
- November 2022: Participated in the Interpretability hackathon organized by Apart Research
- November 2022: Participated in the ML4G, a one-week French AI alignment camp organized by Effisciences
- October 2022: Participated in the AI alignment Hackathon organized by EffiSciences
- 2021-2022: Implemented local playout optimization in a Monte-Carlo tree search for the travelling salesman problem

Programming Projects

- Developed mechanistic interpretability tooling: nnterp, a wrapper around `NNsight` focused on LLMs and tiny-dashboard, a tool to visualize activations of sparse dictionaries
- Early adopter of the mechanistic interpretability library `NNsight`, actively engaging with the community to answer questions and improve it
- Developed a CodinGame multiplayer game
- Ranked in top percentiles in CodinGame multiplayer bot programming contests: top 0.5%, 3% and 7% in 2021-2022
- Proficient in OCaml, Python, Java, PyTorch, and NNsight, with working knowledge of Rust, CUDA, C++ and others

Referees

Neel Nanda

Mechanistic Interpretability Team Lead
DeepMind
neelnanda27[at]gmail[dot]com
www.neelnanda.io

Robert West

Associate Professor and Head of the Data Science Lab
École Polytechnique Fédérale de Lausanne
robert[dot]west[at]epfl.ch
dlab.epfl.ch/people/west

Chris Wendler

Postdoctoral Researcher in the interpretable neural networks lab
Northeastern University
chris[dot]wendler[at]epfl.ch
wendlerc.github.io

Jobst Heitzig

Leader, FutureLab on Game Theory and Networks of Interacting Agents
Potsdam Institute for Climate Impact Research
jobst[dot]heitzig[at]pik-potsdam.de
www.pik-potsdam.de/members/heitzig

Walter Laurito

Research Engineer and Team Lead
Cadenza Lab
lauritowal[at]yahoo[dot]com
www.linkedin.com/in/walter-laurito-951565144

Matthias Fuegger

Head of the Distributed computing group
Formal Methods Laboratory
mfuegger[at]lmf[dot]cnrs.fr
www.lsv.fr/~mfuegger

Charbel-Raphaël Segerie

Executive Director
CeSIA
crsegerie[at]gmail[dot]com
crsegerie.github.io