Davide Ghilardi

@ davide.ghilardi0@gmail.com | LinkedIn | GitHub | Website

Education

University of Milano Bicocca — Milan, Italy
M.Sc. in Data Science; ***GPA: 4.00/4.00*** — *Sep 2022 – Expected Oct 2024*

University of Milano Bicocca — Milan, Italy
B.Sc. in Statistical and Economic Sciences; ***GPA: 3.89/4.00*** — *Sep 2019 – Jul 2022*

Research and Industry Experience

Stanford University — Stanford, California
Visiting Student Researcher — *Jan 2024 – Mar 2024, Full-time*

- Studied and evaluated LLMs to understand their safety behaviors using a linguistic approach under the supervision of Prof. Dan Jurafsky.
- Applied mechanistic interpretability techniques in the context of LLM Safety under the supervision of Federico Bianchi and Mert Yuksekgonul.
- Worked to develop defense techniques against jailbreaking strategies for LLMs in collaboration with Moussa Doumbouya and under the supervision of Prof. Dan Jurafsky and Prof. Christopher Manning.

Consorzio Interuniversitario Nazionale per l'Informatica (CINI) — Milan, Italy
Fellow Researcher — *Sep 2023 – Dec 2023, Full-time*

- Worked on the "Datalake" project in a collaborative effort with the University of Milano Bicocca and under the supervision of Prof. Matteo Palmonari to improve data processing and information extraction of Italian legal investigations.
- Studied and implemented state-of-the-art techniques in entity recognition, extraction, and linking in the context of the Italian legal domain.

University of Milano Bicocca — Milan, Italy
Fellow Researcher — *Apr 2023 – Aug 2023, Full-time*

- Worked on the "Next Generation UPP" project to improve data processing and information extraction applications within the Italian justice system.
- Studied and applied deep learning techniques for information retrieval and knowledge base population in the legal domain under the supervision of Prof. Matteo Palmonari.

Publications

[1] **h4rm3l: A Dynamic Benchmark of Composable Jailbreak Attacks for LLM Safety Assessment**
Moussa Koulako Bala Doumbouya, Ananjan Nandi, Gabriel Poesia, Davide Ghilardi, Anna Goldie, Federico Bianchi, Dan Jurafsky, Christopher D. Manning. *arXiv preprint arXiv:2408:04811*, 2024.

[2] **Accelerating Sparse Autoencoder Training via Layer-Wise Transfer Learning in Large Language Models**
Davide Ghilardi, Federico Belotti, Marco Molinari. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP (EMNLP 2024 Workshop)*, 2024.

[3] **Efficient Training of Sparse Autoencoders for Large Language Models via Layer Groups**
Davide Ghilardi, Federico Belotti, Marco Molinari. *arXiv preprint arXiv:2410.21508 (under review at ICLR 2025)*, 2024.

[4] **Comparing Coarse and Fine-grained Mechanistic Interpretability in Language Models and Getting the Best of Both Worlds**
Nicole Nobili, Davide Ghilardi, Chen Bo Calvin Zhang, Federico Bianchi, Mert Yuksekgonul. *under review at NAACL 2025*, 2024.

Reviewer: ICLR'25, NeurIPS'24.

Achievements and Awards

Neel Nanda's MATS training program participant

- Chosen as one of the 33 participants for Neel Nanda's SERI MATS training program, aimed at advancing research in Mechanistic Interpretability.
- Applied interpretability techniques to analyze the reasoning processes of LLMs through Chain of Thought (CoT), with a particular focus on identifying and addressing unfaithful reasoning patterns[1].

Representative Llama Impact

- Selected, after leading The Good Scientists application for the META Llama Impact Grant, as representative of the Llama Impact Community for Social Impact.

LeadTheFuture Mentee

- Selected as a mentee for the LeadTheFuture organization, a prestigious STEM mentorship program with an acceptance rate below 12%, where I am mentored by top professionals in AI to develop skills and pursue international opportunities.

Volunteering Experience

ML specialist at The Good Scientists: Deployed ML models and NLP techniques to optimize matching processes, delivering tangible benefits to the organization's stakeholders. (Sep 2023 - Present)

Speaker at Datapizza: Collaborated with Datapizza, the most popular Italian media dedicated to Data Science and Artificial Intelligence, to organize and deliver speeches about AI to Italian high-school students. (Mar 2023 – Aug 2023)

[1]You can find our final MATS training program presentation here.