

Education

University of Milano - Bicocca | University of Groningen, CLCG

Milan, Italy / Groningen, Netherlands

Joint Doctorate (Ph.D) in Computer Science - NLP

Nov 2023 – exp. 2026

Research focuses on the use of interpretability as a tool to make generative models safer, more reliable and less toxic in order to extend and improve their real-world applications

University of Milano - Bicocca

Milan, Italy

Master of Science (M.Sc) in Computer Science

Oct 2021 – Oct 2023

Thesis: On the explainability of Large Language Models detoxification

Curriculum: AI & Machine Learning

Final Grade: 110/110, with honors

University of Milano - Bicocca

Milan, Italy

Bachelor of Science (B.Sc) in Computer Science

Oct 2018 – Jul 2021

Thesis: Hate speech detection on social networks using state-of-the-art Natural Language Processing techniques

Final grade: 110/110, with honors

Work Experience

University of Groningen, Computational Linguistic Group

Apr 2023 – Jul 2023

Research Intern

Groningen, Netherlands

- Study, research and development of RLHF/RLAIF and fine-tune algorithms applied to generative language models to modify their behaviour in generating hate speech, effectively automating and controlling the detoxification process and counter-narrative generation;
- Research on the interpretability of the models themselves, analyzing their shift as a result of the post-training procedures adopted.

C.I.N.I. & University of Milano - Bicocca

Jun 2022 – Apr 2023

Assistant Researcher

Milan, Italy

- Research project commissioned by the Italian Ministry of Justice and CINI (Italian Interuniversity Consortium for Computer Science).
- Open Relation Extraction on criminal sentences - Improving state-of-the-art for Italian technical and legal texts understanding using text mining techniques and seq2seq transformer-based Language Models.

TESTUDO

Jun 2022 – Aug 2022

Intern Machine Learning Engineer

Milan, Italy

- Designed and created with University MIND Lab team deep learning models to predict timestamps regarding working hours;
- Worked on non-ideal data to automate the company's recognition system.

University of Milano - Bicocca

Mar 2021 – May 2021

Intern

Milan, Italy

- University internship on Hate Speech detection on Tik-Tok comparing lexicons and using language models to automate content moderation;

Publications

Let the Models Respond: Interpreting Language Model Detoxification Through the Lens of Prompt Dependence

Daniel Scalena, Gabriele Sarti, Malvina Nissim, Elisabetta Fersini

Extended abstract at the Sixth BlackboxNLP Workshop (EMNLP 2023)

MIND at SemEval-2023 Task 11: From Uncertain Predictions to Subjective Disagreement

Giulia Rizzi, Alessandro Astorino, **Daniel Scalena**, Paolo Rosso, Elisabetta Fersini

Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)

Projects, Courseworks and Research works

Let the Models Respond: Interpreting the Detoxification process of LMs

Spring 2023

Research project

[Journal](#)

- Reached enhancement and detoxification of Generative Language Models;
- Aim not to limit model generation with toxic prompts but to foster counter-narrative.
- LMs interpretability techniques for detoxification process evaluation after fine-tuning and/or RLHF;

Reward LM

Spring 2023

Open source Python library

[GitHub](#)

- Python toolkit library enabling Fine-Tuning and Reinforcement Learning from AI Feedback (RLAIF) of Generative Large Language Models;
- Deeply integrated with HuggingFace and Accelerate frameworks;
- A structured methodology for measuring model toxicity is provided, allowing measurement with datasets recognised in the scientific literature.

From Uncertain Predictions to Subjective Disagreement

Spring 2023

Research project & publication

[ACL Anthology](#)

- Participation of the research lab MIND from UniMiB in the SemEval 2023 task related to Learning With Disagreements (Le-Wi-Di);
- Study the identification of the level annotator's agreement/disagreement;
- Explored the correlation between disagreement and uncertainty that a model, based on several linguistic characteristics, could have on the prediction of a given gold label.

-> **Other projects and Open Source contributions** are available on my personal [GitHub profile](#)

Language Skills

Italian Native Language

English Badge Bbetween Foreign Languages – English C1 

French B1

Activities

2024 MSc degree in Data Science at UniMiB

2024 LREC conference, Turin

2023 EMNLP conference, Singapore

2023 European Researchers Night

Teaching NLP course labs

Reviewer

Student Volunteer

GroNLP collaborator, building demo tools on LMs Interpretability