# **Daniel Scalena**

**L** +39 331 \*\*\* \*\*58 (anielsc4.it Milan, IT | Groningen, NL G danielsc4

in daniel-scalena □ scalena99@gmail.com

#### **Education**

## University of Milano - Bicocca | University of Groningen, CLCG

*Nov 2023 – exp. 2026* 

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

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

## University of Milano - Bicocca

Oct 2021 – Oct 2023

Master of Science (M.Sc) in Computer Science, (track: AI & ML)

Thesis: On the explainability of Large Language Models detoxification

Final Grade: 110/110, with honors

### University of Milano - Bicocca

Oct 2018 – Jul 2021

Bachelor of Science (B.Sc) in Computer Science

Thesis: Hate speech detection on social networks using state-of-the-art NLP techniques

*Final grade:* 110/110, with honors

## Work Experience

**Research Intern** *Apr 2023 – Jul 2023* 

University of Groningen, Computational Linguistic Group

Aligned LMs to counter hate speech narratives using RLAIF, reaching up to 50% in safety improvements;

• Analyzed and interpreted model behavior shifts after applying post-training algorithms.

**Assistant Researcher** 

Jun 2022 – Apr 2023

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

Devised Open Relation Extraction for Italian legal texts, improving previous accuracy by 30%;

- Reduced manual legal document review using ad hoc models;
- Project funded by the Italian Ministry of Justice and Interuniversity Consortium for Computer Science.

## **ML Engineer (contractor)**

Jun 2022 – Aug 2022

**TESTUDO** 

- Designed and trained deep learning models to predict working hours timestamps.
- Fully automated the company's recognition system.

## **Scholarships and Awards**

2023 Google and Apple - BlackBoxNLP (EMNLP 2023) travel award Full-Fee M.Sc. Scholarship due to merit and financial scores 2021-2023 2018-2021 Full-Fee B.Sc. Scholarship due to merit and financial scores

## **Publications**

## **𝒇** CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian

Giuseppe Attanasio ... Daniel Scalena

Proceedings at the Tenth Italian Conference on Computational Linguistics, CLIC-it 2024

Multi-property Steering of Large Language Models with Dynamic Activation Composition Daniel Scalena, Gabriele Sarti, Malvina Nissim

Proceedings at the Seventh BlackBoxNLP (at EMNLP 2024)

# **𝚱** A gentle push *funziona benissimo*: making instructed models in Italian via contrastive activation steering

Daniel Scalena, Elisabetta Fersini, Malvina Nissim

Proceedings at the Tenth Italian Conference on Computational Linguistics, CLIC-it 2024

# **𝚱** 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 (at EMNLP 2023)

## **𝒇** 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)

### Personal projects

**Reward LM**Open source Python library
Spring 2023
GitHub

- Python toolkit library for Fine-Tuning and Reinforcement Learning from AI Feedback of Generative LMs;
- Tools to interpret and explain the differences in toxicity between the original models and those trained with different training techniques.

-> Other projects and Open Source contributions are available on my personal GitHub profile

#### **Activities**

**Co-organiser** Challenge the Abilities of LAnguage Models in ITAlian (CALAMITA) (2024)

**Teaching** NLP course labs, MSc Data Science @ UniMiB (2024)

**Reviewer** ACL (2025), LREC-Coling (2024), NAACL (2024\*), Information Processing and Man-

agement (2024), CliC-IT (2023, 2024), ICML Mechanistic Interpretability Workshop

(2024\*) – (\* secondary reviewer)

**Volunteer** EMNLP (2023), CliC-IT (2021)

**Misc.** European Researchers Night GroNLP collaborator (2023)

#### Language Skills

**Italian** Native Language

English Badge Bbetween Foreign Languages – English C1 🔗

French B1