Daniel Scalena

 Italy | Netherlands   ✉ scalena99@gmail.com   📞 +39 331*****58   🔗 danielsc4.it

Education

Double Doctorate (Ph.D) in Computer Science - NLP   *Nov 2023 – exp. Q4 2026*
University of Milano - Bicocca | University of Groningen (CLCG)

Researching interpretability to improve the safety, reliability, and real-world applicability of generative models.

Mediterranean Machine Learning (M2L) Summer School   *Sep 2024*
Milano, Italy

1-week full-time lectures and practical sessions led primarily by Google DeepMind scientists; presented my paper.

M.Sc in Computer Science, ML track   *Oct 2021 – Oct 2023*
University of Milano - Bicocca

- Final Grade: 110/110 with honors, (GPA: 29.7/30)
- Thesis: On the explainability of Large Language Models detoxification.

B.Sc in Computer Science   *Oct 2018 – Jul 2021*
University of Milano - Bicocca

- Final Grade: 110/110 with honors, (GPA: 28.3/30)
- Thesis: Hate speech detection on social networks using state-of-the-art NLP techniques.

Experience

University of Groningen, Computational Linguistic Group   *Apr 2023 – Jul 2023*
Research Intern

- 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.

C.I.N.I. & University of Milano - Bicocca   *Jun 2022 - Apr 2023*
Assistant Researcher

- 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 *Inter-university Consortium for Computer Science.*

TESTUDO s.r.l.   *Jun 2022 - Aug 2022*
ML Engineer (contractor)

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

Main publications

⬈ Steering Large Language Models for Machine Translation Personalization
Daniel Scalena*, Gabriele Sarti*, Arianna Bisazza, Malvina Nissim, Elisabetta Fersini
ArXiv preprint - under review, EMNLP 2025

⬈ 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)

⬈ A gentle push funziona benissimo: making instructed models in Italian via contrastive activation steering
Daniel Scalena, Gabriele Sarti, Malvina Nissim, Elisabetta Fersini
Proceedings at the Tenth Italian Conference on Computational Linguistics, CLIC-it 2024

[↗] **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)

Full list of publications [↗]

Scholarships and Awards

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

Projects

Reward LM *DanielSc4/RewardLM*[↗]
  - Python toolkit library for Fine-Tuning and Reinforcement Learning from AI Feedback of Generative LMs;
  - Developed tools to interpret and compare toxicity differences between original and fine-tuned models.

Activities

Co-organiser	Challenge the Abilities of LAnguage Models in ITAlian (CALAMITA) *2024*
Teaching	NLP course labs, MSc Data Science @ UniMiB (*2024 - 2025*)
Reviewer	EMNLP *2025*, ACL *2025*, LREC-Coling *2024*, NAACL *2024**, Information Processing and Management *2024 - 2025*, CliC-IT *2023 - 2024*, ICML Mechanistic Interpretability Workshop *2024** (*secondary reviewer)
Volunteer	CliC-IT *2021*, EMNLP *2023*
Misc.	European Researchers Night GroNLP collaborator *2023*

Languages

Italian	Native Language
English	C1 - B-between Foreign Languages [↗]
French	B1