Pietro **Lesci**

Email: pietrolesci@outlook.com **Personal Page:** pietrolesci.github.io
Google Scholar: pietrolesci **LinkedIn:** linkedin.com/pietrolesci
Github: github.com/pietrolesci **Twitter-X:** @pietro_lesci

PhD student in NLP at Cambridge University with a background in econometrics. 3+ years experience across research labs, consulting firms, and international institutions training custom models and developing data science solutions. Expert in developing causal methods to study the effect of training data on models' behaviours.

PROFESSIONAL EXPERIENCE

PhD Student @ University of Cambridge
October 2021 – present | Cambridge, UK
I work with Prof Andreas Vlachos on causal methods to study the effect of training data on models' behaviours, including memorisation, generalisation, and tokenisation. My work includes:

- Large-scale data processing pipelines
- Implementation and pre-training of LLMs
- Non-standard training loops (e.g., active training)

Applied Scientist Intern @ Amazon AWS AI Labs
September 2022 – January 2023 | Barcelona, ES
I worked with Lluis Marquez on efficient dialogue state tracking with language models (published at ACL 2023). My work included:

- Developing methods to scale models to long conversations
- Fine-tuning language models for dialogue tasks
- Curating datasets for model evaluation

Senior Associate, Data Science @ Bain & Company
March 2020 – October 2021 | Milan, IT
I solved complex business problems across various industries (e.g., mining, energy) using data science. My work included:

- Generating and prioritising solutions (*"80/20"* approach)
- Presenting solutions and their tradeoffs to clients
- Coordinating the work of junior consultants on the team

Research Assistant @ Bocconi University BIDSA Center
August 2019 – March 2020 | Milan, IT
I worked with Prof Dirk Hovy on the identification of online abuse. My work included developing the *Wordify* and *MACE* web-apps.

Data Science Intern @ European Central Bank
February 2018 – November 2018 | Frankfurt am Main, DE
I worked on enriching internal databases. My work included:

- Data ingestion, integration, and quality assurance
- Developing custom data deduplication pipelines
- Reporting to experts from the National Central Banks

SELECTED FIRST-AUTHOR PUBLICATIONS

[1] *PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs*. ICLR 2025.

[2] *Causal Estimation of Memorisation Profiles*. ACL 2024 (Best Paper Award).

[3] *AnchorAL: Computationally Efficient Active Learning for Large and Imbalanced Datasets*. NAACL 2024.

AWARDS & GRANTS

Best Paper Award at ACL 2024
Our paper was among the 5 winners (0.25% of all accepted papers).

Translated Imminent Research Grant
Our project obtained $20k in funding.

EDUCATION

MSc in Economic and Social Sciences @ Bocconi University
September 2016 – March 2019 | Milan, IT
Bayesian Statistics and Econometrics. Dissertation: *"Deep Learning: A Statistical Perspective"*. Grade: 104/110

BSc in Economics and Management @ Università Cattolica del Sacro Cuore
September 2013 – September 2016 | Milan, IT
Macroeconomics and Agent-Based Models. Dissertation: *"Market Sentiment and Monetary Policy"*. Grade: 110/110 cum laude

Scientific High School Diploma
September 2008 – July 2013 | Castrovillari, IT
Grade: 95/100

ACADEMIC ROLES

Teaching Assistant and Thesis Supervisor
A.Y. 2022, 2023, 2024 | Cambridge, UK
Lead practical and marking sessions (course "Machine Learning and Real-world Data", 300 students). Co-supervised MSc thesis projects.

Conference Reviewer
2019 – present
Reviewed for top-ranked international conferences (ACL, NAACL, EMNLP, ICLR).

SKILLS

Natural & Formal Languages
Italian (native), English (fluent), and French (basic). Python, R, Stata, SQL, LaTeX, Julia, MATLAB, and Rust.

EXTRACURRICULAR

Music Education
Classical piano and jazz bass guitar.