Daniele Potertì

PhD candidate in LLM interpretability & evaluation

daniele@poterti.com
daniele.poterti.com
Google Scholar
danielepoterti LinkedIn

Italian (EU) citizen | PhD defence expected March 2028

RESEARCH INTERESTS

Interpretability and evaluation of large language models: activation steering and role/persona directions, sparse-autoencoder feature evaluation, theory-driven interpretability of deceptive behaviour, and robustness of LLM reasoning to malicious inputs.

EDUCATION

Ph.D. in Economics, Statistics and Data Science (ECOSTATDATA) 2023 – 2028 (exp.)
University of Milano-Bicocca Milan, Italy

- *Thesis (working title):* Evaluating and Understanding Large Language Models
- *Supervisor:* Prof. Fabio Mercorio; *co-supervisor:* Prof. Erik Cambria (NTU Singapore)

Visiting Ph.D. Student May 2026 – Dec 2026
Nanyang Technological University (NTU), host: Prof. Erik Cambria Singapore

- Theory-driven (deductive) interpretability of deception: testing whether LLM internals encode the dimensions predicted by Information Manipulation Theory. Paper in preparation.

M.Sc. in Data Science 2021 – 2023
University of Milano-Bicocca Milan, Italy

- *Thesis:* Bridging Causal Theory and LLMs: A User-Centric Approach to Causal Graph Generation (with Intesa Sanpaolo)

B.Sc. in Computer Science 2018 – 2021
University of Milano-Bicocca Milan, Italy

EXPERIENCE

Research Intern Mar 2023 – Oct 2023
Intesa Sanpaolo, Data Science & AI Office Milan, Italy

- LLM-assisted causal-graph construction with domain-expert validation; led to the AI*IA 2023 workshop paper [9]; follow-up work on LLM-generated explanations in banking [8].

PUBLICATIONS

My group lists authors alphabetically; † marks papers on which I was the lead author.

Interpretability and safety

[1] † F. Mercorio, **D. Potertì**, A. Seveso. *Diagnosing LLM Fragility: An Empirical Study on Reasoning Entropy and Malicious Outputs.* **EMNLP 2026 (Main)**.
Contribution: equal contribution, authors listed alphabetically; corresponding author. Led the project: research idea, experimental design, implementation and writing.

[2] † **D. Potertì**, A. Seveso, F. Mercorio. *Can Role Vectors Affect LLM Behaviour?* **Findings of EMNLP 2025**.
Contribution: led the project: research idea, role-vector design, steering experiments and writing.

[3] † F. Mercorio, F. Pallucchini, **D. Potertì**, A. Serino, A. Seveso. *SFAL: Semantic-Functional Alignment Scores for Distributional Evaluation of Auto-Interpretability in Sparse Autoencoders.* **EMNLP 2025 (Industry)**.
Contribution: co-main contributor with A. Serino: proposed the core idea of representing SAE features through their co-occurrence statistics, and co-led the technical implementation.

LLM evaluation and benchmarks

[4] A. Seveso, **D. Potertì**, E. Federici, M. Mezzanzanica, F. Mercorio. *ITALIC: An Italian Culture-Aware Natural Language Benchmark.* **NAACL 2025 (Long)**.
Contribution: main technical contributor.

[5] F. Mercorio, M. Mezzanzanica, **D. Potertì**, A. Serino, A. Seveso. *A Benchmark to Evaluate LLMs' Proficiency on Italian Student Competencies.* **ECML PKDD 2025**.
Contribution: technical contribution (implementation and experiments).

[6] F. Mercorio, **D. Potertì**, A. Serino, A. Seveso. *BEEP – BEst DrivEr's License Performer: A CALAMITA Challenge.* **CLiC-it 2024**.

[7] M. Nissim, D. Croce, V. Patti, et al. (incl. **D. Potertì**). *Challenging the Abilities of Large Language Models in Italian: a Community Initiative.* **IJCoL 2026**.

Explainability and causality

[8] A. Castelnovo, R. Depalmas, F. Mercorio, N. Mombelli, **D. Potertì**, A. Serino, A. Seveso, S. Sorrentino, L. Viola. *Augmenting XAI with LLMs: A Case Study in Banking Marketing Recommendation.* **xAI 2024**, Springer CCIS.
Contribution: technical contribution (implementation and experiments).

[9] A. Castelnovo, R. Crupi, F. Mercorio, M. Mezzanzanica, **D. Potertì**, D. Regoli. *Marrying LLMs with Domain Expert Validation for Causal Graph Generation.* **AIABI Workshop @ AI*IA 2023**.

[10] **D. Potertì**. *LLMs in the Neurosymbolic Cycle.* **Doctoral Consortium @ ECAI 2024**.

[11] **D. Potertì**. *Bridging Causal Theory and LLMs: A User-Centric Approach to Causal Graph Generation.* **Doctoral Consortium @ AI*IA 2023**.

COMPUTE AND TECHNICAL SKILLS

HPC	Multi-node, multi-GPU training and inference on the Leonardo supercomputer (CINECA / EuroHPC) through three ISCRA-C grants (PI on the most recent). Local: RTX 5090 (32 GB).
ML	PyTorch; Hugging Face Transformers, Accelerate, Datasets; vLLM; TransformerLens and custom forward hooks for activation extraction and steering; Inspect AI for evaluations; scikit-learn.
Programming	Python, Go, Java, C++; Docker, Slurm, Linux.

SERVICE AND LANGUAGES

Reviewing	ACL Rolling Review (ARR); external reviewer for ECML PKDD 2024.
Languages	Italian (native); English (C1, CEFR).