Francesco Ortu *PhD Student*

francescortu.github.io   francescortu   francescortu   francesco.ortu@phd.units.it   Trieste,Italy

Education

Ph.D. Student — Oct 2024 – present
University of Trieste, AREA Science Park — Trieste, Italy

- *Research Interest:* AI Safety, LLM/VLM Interpretability, NLP4Good.
- *Expected graduation:* Sep 2027

MSc: Data Science and Scientific Computing — Mar 2024
Joint program between University of Trieste, SISSA, ICTP — Trieste, Italy

- *Track*: Foundation of AI and ML; all courses taught and assessed in English.
- *Relevant courses*: Deep Learning, Reinforcement Learning, Probabilistic ML
- *GPA:* 110/110 with Honors.

BSc: Mathematics — Dec 2020
Sapienza University of Rome — Rome, Italy

Publications

The Narrow Gate: Localized Image-Text Communication in Vision-Language Models — 2024
Alessandro P. Serra, **Francesco Ortu*,** Emanuele Panizon*, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, Aberto Cazzaniga*; **Under Review**

Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals — 2024
Francesco Ortu*, *Zhijing Jin*, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, Bernhard Schölkopf;* **ACL 2024 Main**

Language Model Alignment in Multilingual Trolley Problems — 2024
*Zhijing Jin, Sydney Levine, Max Kleiman-Weiner, Giorgio Piatti, Jiarui Liu, Fernando Gonzalez Adauto, **Francesco Ortu**, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, Bernhard Schölkopf;* **Pluralistic Alignment @ NeurIPS 2024**

Experience

Research Fellow — May 2024 – present
AREA Science Park — Trieste, Italy

- Researching in AI Safety and NLP4Good supervised by Alberto Cazzaniga.

Research Intern — Sep 2023 – Jan 2024
Max-Planck Institute for Intelligent Systems — Tübingen, Germany

- Supervised by Prof. Bernhard Schölkopf and Zhijing Jin.
- Worked on AI Safety and interpretability of LLM.
- Using causal techniques (mechanistic interpretability) to understand the interplay between fact recall and contextual learning in LLMs.
- Paper was accepted to ACL2024 main conference

Machine Learning Engineer Intern — Feb 2023 – Apr 2023
PLUS — Trieste, Italy

- Conducted research as NLP engineer in AI team.
- Develop proof of concept for advanced information retrieval in business applications.
- Created RAG system prototype using Llama-based LLMs, focusing on retrieval optimization.
- Demonstrated potential for significant efficiency gains in knowledge-driven applications.

Skills

Python (*PyTorch, HuggingFace Transformers, Scikit-Learn, Pandas*) | **C/C++** (*OpenMP, MPI*) | **R** | **SQL** | **Bash** | **Git** | **Linux**

Teaching

Teaching Assistant - Natural Language Processing 2024

University of Trieste

- Crafted and taught jupyter notebooks on NLP pipelines, transformers and LLMs. Master level course.

Scholarships & Honors

Best Thesis in AI 2024

University of Trieste

- Recognized for the best thesis in the Data Science program since its establishment in 2018.

Best student of 2021 2022

University of Trieste

- Scholarship for being among the top students of the 2021 cohort.