Jacopo Zacchigna

Trieste, Italy | jaczac2002@gmail.com | jac-zac.github.io | github.com/Jac-Zac
linkedin.com/in/jacopo-zacchigna | Google Scholar

Summary

Final-year MSc student in Data Science & AI, University of Trieste, researching mechanistic interpretability of LLMs with interests spanning Mixture-of-Experts models and representation geometry. Former SPAR Research Fellow.

Research Experience

SPAR — Supervised Program for Alignment Research, *Research Fellow* Feb 2026 – May 2026

Characterizing and steering implicit personalization in LLMs.

- Extracted and released persona vectors for Gemma-2 (9B/27B) and Llama-3.1 (70B/405B), enabling steering and intervention on learned user representations.
- Developed `persona-data`, a Python package standardizing access to 1k synthetic personas.
- Built an interactive web UI for real-time extraction, analysis, and probing of implicit persona representations.

Mechanistic Interpretability of Mixture-of-Experts LLMs, *Independent research* 2026 – present

- Adapting HeadPursuit / SOMP to identify expert specialization (e.g., toxicity) in OLMoE and GPT-OSS.
- Investigating expert editing and steering interventions versus knockout-style ablations.

Selected Projects

BayesianFlow — Uncertainty Estimation in Generative Models, code Jun – Jul 2025

- Extended BayesDiff to flow matching, achieving 5× faster generation than DDIM at comparable quality.
- Integrated Last-Layer Laplace Approximation for pixel-wise uncertainty (*interpretable* confidence maps)

Causal Multi-Head Self-Attention Kernels, code 2025

- Implemented causal MHSA in CUDA, OpenMP, and SIMD; achieved 1.09× speedup over naive PyTorch on A100, and built towards a FlashAttention-style tiled kernel.

Publications

Zacchigna, J. & Liu, W. (co-first authors), et al. *Application of Computer Vision to the Automated Extraction of Metadata from Natural History Specimen Labels: A Case Study on Herbarium Specimens.* Plants (MDPI) 2026

Education

University of Trieste, MSc, Data Science & Artificial Intelligence – Trieste, Italy 2024 – 2026
Relevant coursework: Theory of Deep Neural Networks, Explainable AI, Probabilistic ML, Advanced Deep Learning, NLP, Reinforcement Learning, Unsupervised Learning.

University of Trieste, BSc, AI & Data Analytics – Trieste, Italy 2021 – 2024

Experience

Obloo, AI Consultant – Trieste, Italy May 2024 – Jan 2025

- Designed retrieval-augmented generation (RAG) systems for knowledge retrieval over the company data lake
- Mentored 5 interns to independent project delivery; led ML prototypes adopted in two company-wide products

Skills

Technical: Python, C/C++, CUDA, PyTorch, NNsight/nnterp, Hugging Face, NumPy, R, Docker, Linux, HPC

Languages: Italian (native), English (C1)