Matteo Nulli

Website, GitHub, Linkedin, Twitter, Blog

Email : matteo.nulli@outlook.com
Mobile : +39-334-248-4311

Research Interests Multimodal Learning, Visual Compositional Reasoning, Inference Optimization

Education

- **University of Amsterdam** — Amsterdam, Netherlands
  *MSc in Artificial Intelligence; GPA: 4.0/4.0 (8.5/10) - **Cum Laude** & ELLIS Honours* — *Sep 2023 - Sep 2025*
  - ◇ **Thesis**: Object-Guided Visual Tokens: Eliciting Compositional Reasoning in Multimodal Language Models. Advised by Prof. Yuki Asano, Ivona Najdenkoska, Mohammad Mahdi Derakhshani.
  - ◇ **ELLIS Honours Student**: Selected as ELLIS MSc Honours Student for 2025.
  - ◇ **Relevant Courses**: Foundation Models, Deep Learning 1 & 2, Computer Vision, Natural Language Processing, Information Retrieval, Machine Learning 1.
- **Università Commerciale Luigi Bocconi** — Milan, Italy
  BSc in Mathematics and Computing Sciences for Artificial Intelligence; GPA: 3.6/4.0 (99/110) — *Sep 2020 - July 2023*
  - ◇ **Thesis**: **Generative** Adversarial Networks and Recurrent Neural Networks for time-series asset price prediction. Advised by Prof. Claudio Tebaldi.
  - ◇ **Relevant Courses**: Machine Learning, Mathematical Modelling for Finance, Mathematical Analysis 1,2 & 3, Physics 1 & 2, Statistical and Quantum Physics, Optimization Algorithms, Programming.
- **University of Sydney** — Sydney, Australia
  Exchange Semester in Applied Mathematics and Computing Sciences; GPA: 3.6/4.0 — *Feb 2023 - July 2023*
  - ◇ **Scholarship**: Selected by merit and received a full ride scholarship of $20.000.
  - ◇ **Relevant Courses**: Stochastic Processes (Adv), Big Data and Data Diversity (Adv), Deep Learning

Publications & Preprints

- ○ Meeting SLOs, Slashing Hours: Automated Enterprise LLM Optimization with OptiKIT, **MLSys** Industry Track, 2026.
- ○ Adapting Vision-Language Models for E-commerce Understanding at Scale, **EACL** Industry Track, 2026.
- ○ Object-Guided Visual Tokens: Eliciting Compositional Reasoning in Multimodal Language Models, **EurIPS** 2025 Principles of Generative Modeling, Copenhagen, Denmark.
- ○ In-Context Learning Improves Compositional Understanding of Vision-Language Models, **ICML** 2024 Foundation Models in the Wild, Vienna, Austria.
- ○ Dynamic Vocabulary Pruning in Early-Exit LLMs, **NeurIPS** ENLSP 2024, Vancouver, Canada.
- ○ 'Explaining RL Decisions with Trajectories': A Reproducibility Study. Transactions of Machine Learning Research - **TMLR**, 2024, Vancouver, Canada.

Experience

- **Applied Researcher** — Amsterdam, Netherlands
  eBay, Foundation Models Team — *Aug 2025 - Current*
  - ◇ **Focus**: Research on Inference Optimization and Multimodal Search-Reasoning models (MLSys Submission).
- **Applied Research Intern** — Amsterdam, Netherlands
  eBay, Foundation Models Team, advised by Prof. Cees Snoek — *Jul 2024 - Jul 2025*
  - ◇ **Focus**: Research on **Multimodal Learning**. Training VLMs for e-Commerce tasks while conducting theoretical research on architecture. Developed eBay's **Vision-Language Models**, executing comprehensive domain-adaptation, training the model from scratch on a blend of open-source and proprietary data. Working with Hadi Hashemi and Vladimir Orshulevich.
- **Research Intern** — Nuremberg, Germany
  Fundamental AI Lab, University of Technology Nuremberg, advised by Prof. Yuki Asano — *Mar 2025 - Jul 2025*
  - ◇ **Focus**: Researching **Compositional Reasoning** and Visual Grounding in Multimodal Learning (EurIPS PriGM).

- **Co-Founder** Milan, Italy
  BAINSA *Jan 2022 - July 2023*
  - ◇ **Focus**: Founded first Artificial Intelligence association at Bocconi. Spreading awareness & perception on AI's applications through events held inside and outside the university.
  - ◇ **Partners**: Main Partners included Bending Spoons, Vedrai and Insitute Europia.
- **Research Intern** Milan, Italy
  Aindo, advised by Sebastiano Saccani *Jun 2022 - Sep 2022*
  - ◇ **Focus**: Built and deployed deep learning models for synthetic data generation, specializing in VAEs.

Scholarships & Awards

- **Award**: *111 Student List 2025 by Nova*
  Selected as one of the 10 most promising Italian students in Mathematics & Data Analytics, part of the Nova 111 Student List 2025.
- **Scholarship**: *ELLIS Honours Student by ELLIS Unit Amsterdam*
  ELLIS Honours Student scholarship (top 5% of students) of $3.000 supporting research visits, 2025.
  Here is a video of my presentation
- **Scholarship**: Full ride by Università Commerciale Luigi Bocconi @ *University of Sydney*
  Received a full ride scholarship of $20.000 to attend semester abroad at University of Sydney, 2023.
- **Award**: Competition by *Università Commerciale Luigi Bocconi*, presentation @ *University of Oxford*,
  Won a ML competition to analyze Hypoxia in **breast cancer cells**. Awarded founding to present our research at the **University of Oxford**, Oncology Department. Advised by prof. Francesca Buffa, 2022.
- **Scholarship**: Early Academic Excellence by *Mario Negri Foundation*
  Merit scholarship of $1500 for outstanding high-school performance in Mathematics and Physics, 2019.

Blogposts & Projects

- **Machine Learning for Breast Cancer Cells analysis, University of Oxford, 2022**: Utilized Unsupervised and Supervised ML methods (Tree based methods and Deep Neural Network) to analyse breast cancer cells and capture interactions between them. Detected with a success rate of 95% Hypoxic vs Normoxic cells. Won the competition among peers and presented our findings at the University of Oxford Oncology Department. Advised by Prof. Francesca Buffa.
- **Optimizing Predictions: Vocabulary Reduction and Contrastive Decoding in LLMs, University of Amsterdam, 2024**: In this blogpost we investigate early-exiting mechanisms for LLMs as a way to reduce inference cost while preserving accuracy, highlighting calibration issues in non–fine-tuned models and proposing heuristics to mitigate them. We introduce vocabulary pruning to speed up inference with minimal performance loss and complement it with within-model contrastive decoding to maintain confidence. Experiments on summarization and question answering show that combining these techniques yields a Pareto improvement in both computational efficiency and model performance.
- **Perception, Localization, Planning and Control on RAE Robots, University of Amsterdam, 2024**: Built an integrated perception–localization–planning–control pipeline on a RAE robot. We go over Camera Calibration, Line Following, Localization, Curling Match playing (see Video 1) and Mapping, Planning and Control, to allow our RAE Robot to freely move across our environment.
- **Model compression for Machine Translation, University of Amsterdam, 2024**: This work studies compressing ALMA, a multilingual machine-translation LLM, to reduce its high inference cost while preserving translation quality. It evaluates several quantization and pruning methods on ALMA-7B, analyzing trade-offs between quality, memory, and speed, and shows that combining techniques like Wanda and GPTQ can achieve up to 3.5× memory savings with limited performance loss.

Skills

Programming Languages: Python, R, SQL, LaTeX, C (Beginner)
Libraries: Pytorch, OpenCV, Transformers, SciPy, Pandas, NumPy, Matplotlib, Scikit Learn, CLIP, ...
Languages: Italian (Native), English (Fluent), Spanish (Fluent)

Volunteering

- **Machine Learning Engineer**: *BSI Bocconi - Build Sustainable Innovation, 2021-2023.*
  Implemented ML & Statistical based solutions for Companies. Applied Data analysis techniques to costumer provided datasets.
- **Peer Mentor and Ambassador**: *Università Commerciale Luigi Bocconi, 2021-2023.*
  Mentored international students upon their arrival on campus.