Alessio Devoto

devoto.alessio@gmail.com
alessiodevoto.io
X/devoto_alessio

Experience

Applied Researcher | NVIDIA Feb 2026 – Present
Applied Agent Research & Kaggle GrandMasters Teams.

Intern | NVIDIA Jun 2025 – Oct 2025
Worked on efficiency for LLMs and KVPress library.

Teaching Assistant | Sapienza University Sep 2023 – Nov 2025
Led hands-on PyTorch tutorials and project supervision for 120+ MSc students.

AI Lecturer Assistant | Deepers Sep 2023 – Nov 2025
Delivered high-level technical sessions for executives Deepers bootcamp.

Freelance Developer Jan 2022 – Nov 2025
Developed and deployed on-premise LLM-based and speech-to-text applications.

ICF Trainee Coach | ICF Feb 2020 – Present
Training to become a life & business coach (30+ hours experience as individual coach).

Research Internship | ISPAMM Lab Jan 2022 – Nov 2022
Developed models for explainable High Energy Physics in collaboration with CERN.

Tutor Sep 2016 – Present
Tutor for 40+ university/high school students (Latin, Ancient Greek, Maths).

Education

PhD in Data Science Nov 2022 – Jan 2026
La Sapienza, University of Rome.
Focus on Efficient and Adaptive neural networks and Explainability for AI models.
Supervisor: Prof. Simone Scardapane.

Visiting Researcher Mar 2024 – Jul 2024
The University of Edinburgh.
Focus on NLP with emphasis on efficient inference and explainability.
Supervisor: Prof. Pasquale Minervini

Master's Degree in Computer Engineering Sep 2019 – Jan 2022
La Sapienza, University of Rome – Final mark: 110/110 cum Laude.

Visiting Student Feb 2021 – Jul 2021
Universidad Politecnica de Valencia, Spain.

Bachelor's Degree in Control and Computer Engineering Sep 2016 – Oct 2019
La Sapienza University of Rome – Final mark: 110/110 cum Laude.

High School Diploma Feb 2012 – Jul 2016
Humanities and Ancient Languages (Latin, Ancient Greek) – Final mark: 100/100.

Blog

I maintain a small blog where I share code tutorials and insights on various deep learning topics, like implementing a *"ViT from scratch in pure JAX"* or *"Logitlens from scratch without interpretability libraries"*. Visit my blog here: https://alessiodevoto.github.io/blog.

Research Projects

Explainability for High Energy Physics (with CERN, University of Liverpool) Feb 2023 – Jan 2026
Developed explainability methods for AI models (mainly GNNs) for Science Discovery.
MUCCA Project Website

Next Generation 6G communications. Mar 2023 – Jan 2025
Designed adaptive neural networks for next-gen 6G goal-oriented communication pipelines.
6G-GOALS Website

Selected Publications

A more comprehensive list is available on my Google Scholar profile

A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression. A. Devoto* , Y. Zhao*, S. Scardapane, and P. Minervini. *Empirical Findings in Natural Language Processing (EMNLP)*, 2024. arXiv:2406.11430

Expected Attention: Leveraging Future Queries Distribution for KV Cache Compression . A. Devoto , M. Jeblik, S. Jegou, *Preprint* arXiv:2510.00636

Adaptive Computation Modules: Granular Conditional Computation For Efficient Inference. B. Wójcik, **A. Devoto** , K. Pustelnik, P. Minervini, and S. Scardapane. *Proceeding of 39-th the AAAI Conference on Artificial Intelligence (AAAI)*, 2025. arXiv:2312.10193

Q-Filters: Leveraging QK Geometry for Efficient KV Cache Compression. Nathan Godey, **A. Devoto***, Yu Zhao, Simone Scardapane, Pasquale Minervini, Éric de la Clergerie, Benoît Sagot. *SLLM workshop @ ICLR*, 2025. arXiv:2503.02812

Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering. Y. Zhao, **A. Devoto** , G. Hong, X. Du, A. P. Gema, H. Wang, K.-F. Wong, and P. Minervini. *Nations of the Americas Chapter of the ACL (NAACL)*, 2025. arXiv:2410.15999

Adaptive Layer Selection for Efficient Vision Transformer Fine-Tuning. A. Devoto , F. Alvetreti, J. Pomponi, P. Di Lorenzo, P. Minervini, and S. Scardapane. *Neurocomputing, vol. 654*, 2024. arXiv:2408.08670

Analysing the Residual Stream of Language Models Under Knowledge Conflicts. Y. Zhao, X. Du, G. Hong, A. P. Gema, **A. Devoto** , H. Wang, X. He, K.-F. Wong, and P. Minervini. *Foundation Model Interventions Workshop (MINT) NeurIPS*, 2024 arXiv:2410.16090

Are We Done with MMLU? A. P. Gema, J. O. J. Leang, G. Hong, **A. Devoto** , A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, and M. R. G. Madani. *Nations of the Americas Chapter of the ACL (NAACL)*, 2025. arXiv:2406.04127

Adaptive Semantic Token Selection for AI-native Goal-oriented Communications. A. Devoto , S. Petruzzi, J. Pomponi, P. Di Lorenzo, and S. Scardapane. *Global Communications Conference (GlobeComm)*, 2024 arXiv:2405.02330

Conditional computation in neural networks: principles and research trends. S. Scardapane, A. Baiocchi, **A. Devoto** , V. Marsocci, P. Minervini, and J. Pomponi. *Artificial Intelligence*, 2024. arXiv:2403.07965

Cascaded Scaling Classifier: class incremental learning with probability scaling. J. Pomponi, **A. Devoto** , and S. Scardapane. *Neurocomputing, vol. 460*, 2024. arXiv:2402.01262

Technical Skills

Deep Learning Frameworks: PyTorch, JAX, Hugging Face Transformers

Programming Languages: Python, C, Java

Development Tools: Git, Docker, Unix/Linux

Research Areas: Adaptive & Dynamic Neural Networks, Efficient Inference & Training, AI Interpretability

Web Development: HTML, JavaScript, CSS

Languages

Italian: Native
English: C2
Spanish: C1
Portuguese: B2 & learning