Diego Antognini, PhD

Researcher in Machine Learning & Natural Language Processing

diegoantognini@gmail.com | www.diegoantognini.com | Zürich, Switzerland | Diego Antognini | Diego999 3.0k⋆ | diegoantognini

7 years of research experience in natural language processing, machine learning, and recommendation systems. Focusing on enhancing large multimodal models through iterative feedback and refinement. Worked on aligning large language models,building retrieval-augmented LLM systems, and developing efficient models for low-resource settings. Experienced in designing explainable models that generate personalized and actionable textual explanations. Supervised 70+ B/M.Sc. projects.

Skills

Research Interests	Generative AI, LLM alignment, multimodal, iterative refinement, efficient ML, NLP, conversational recommendation.
Program Committee	NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, EACL, SIGIR, RecSys. Journals: ACL Rolling Review, ACM Computing.
Languages & Libraries	Efficient: Python, PyTorch, Tensorflow, transformers, ONNX, Spark, Bash, SQL. Prior Experience: C++, CUDA, Java.
Technologies	GNU/Linux, Git, Poetry, Docker, Kubernetes, Openshift, API design, Redis, Elasticsearch, Milvus vector database.

Experience

Research Engineer — **Google Research**
Jan. 2024 - present — *Zürich, Switzerland*

- Advancing multimodal generative AI by focusing on large multimodal models and enhancing them through iterative feedback and refinement.

Research Scientist — **IBM Research**
May 2022 - Jan. 2024 — *Zürich, Switzerland*

- Publications: **6 papers in AI & ML leading venues**: 3 conference, 1 workshop, 2 demo, 1 under review at TMLR. Patents: **5 patents that are filed**.
- Implemented new methods for aligning large language models to convert natural language questions into SQL queries for massive databases.
- Designed methods to adapt and personalize LLMs to users, using parameter-efficient fine-tuning methods. To be integrated into *IBM Watsonx.ai*.
- Developed tiny, low-latency models with high performance and throughput. Created a term extractor for technical domains and reduced latency by 10x on CPU while performing similarly to BERT. Built a term encoder matching sentence encoders in quality, yet 5x smaller and 10x faster.
- Deployed models of 1MB and 2ms latency used in *IBM Deep Search* to extract terms in real time from scientific documents and patents.
- Built a distributed system to generate QA pairs from large corpora using LLMs and a retrieval-augmented LLM system to answer users' questions.

Module Head, Lecturer, and Supervisor for M.Sc. Theses in NLP — **Lucerne University of Applied Sciences**
Feb. 2022 - present — *Lucerne, Switzerland*

- Designing and teaching the course of computational language technologies and deep learning for NLP to 160+ M.Sc. students.
- Supervised 10 M.Sc. theses in NLP with companies in the areas of medicine, law, politics, insurances, banks, media, and data visualization.

Consultant and Expert for B.Sc. and M.Eng. Theses in ML — **HE-ARC – University of Applied Sciences**
June 2015 - present — *Neuchâtel, Switzerland*

- Giving talks on a wide range of deep learning topics and offering machine learning consulting services for applied research in industrial projects.
- Assessed 30+ B.Sc./M.Eng. theses in the areas of autonomous drones & driving, algorithmic optimization with GPUs, computer vision, and NLP.

Visiting Researcher in Prof. Julian McAuley's ML Lab — **UCSD – University of California San Diego**
Jul. 2021 - Nov. 2021 — *San Diego, CA, U.S.A.*

- Published an unsupervised critiquing method for generative language models to help users rewrite cooking recipes to satisfy dietary restrictions.

Research and Teaching Assistant — **EPFL – Swiss Federal Institute of Technology in Lausanne**
May 2017 - Mar. 2022 — *Lausanne, Switzerland*

- Assisted in teaching intelligent agents (M.Sc.), introduction to natural language processing (M.Sc.), and artificial intelligence courses (B.Sc.).
- Supervised 30+ B./M.Sc. semester projects & theses. Worked with the data analytics & AI research team in Swisscom (led by Dr. Claudiu Musat).

Education

Ph.D. in Computer Science — **EPFL – Swiss Federal Institute of Technology in Lausanne**
Sep. 2017 - Mar. 2022 — *Lausanne, Switzerland*

- Publications: **15 papers in AI & ML leading venues**: 8 conference, 6 workshop, 1 demo. Advisor: Prof. Boi Faltings, head of the AI laboratory.
- Implemented the **first PyTorch graph attention network, starred and forked on Github 3.3k+** with 10k views per month.
- Thesis: Textual Explanations and Critiques in Recommendation Systems. I solved two challenges: generating textual explanations and making them actionable. My thesis focused on generative AI, explainability, and conversational recommendation. **Fastest to graduate in the AI lab**.

M.Sc. in Computer Science — **EPFL – Swiss Federal Institute of Technology in Lausanne**
Sep. 2014 - Apr. 2017 — *Lausanne, Switzerland*

- Specialization: NLP, AI, ML, and distributed systems (GPA: 5.5/6.0). It includes an extra year of 62 ECTS credits to be accepted in the program.
- Thesis: From Relation Extraction to Knowledge Graphs. Built a model that extracts terms and concepts from large corpora and classifies the semantic relationship between them. It outperformed state-of-the-art models by 0.9 F1-score in the relation-classification task of SemEval-2010.

B.Sc. in Computer Science — **HE-ARC – University of Applied Sciences**
Sep. 2011 - Aug. 2014 — *Neuchâtel, Switzerland*

- Major: software engineering (GPA 5.6/6.0). Thesis: Computing Brain Neuronal Maps. Developed a multi-GPUs algorithm to compute an accurate 3D real-time rendering of the brain's electromagnetic activities. Reduced the computation time from 20h to 700ms (faster by a factor of 100,000).

Publications (selected)

Assistive Recipe Editing through Critiquing — EACL 2023
Diego Antognini, Shuyang Li, Boi Faltings, Julian McAuley

pNLP-Mixer: an Efficient all-MLP Architecture for Language — ACL 2023
Francesco Fusco, Damian Pascual, Peter Staar, Diego Antognini

Extracting Text Representations for Terms and Phrases in Technical Domains — ACL 2023
Francesco Fusco* and Diego Antognini* (equal contribution)

Unsupervised Term Extraction for Highly Technical Domains — EMNLP 2022
Francesco Fusco, Peter Staar, Diego Antognini

Fast Critiquing with Self-Supervision for VAE-based Recommender Systems — RecSys 2021
Diego Antognini and Boi Faltings

Interacting with Explanations through Critiquing — IJCAI 2021
Diego Antognini, Claudiu Musat, Boi Faltings

Rationalization through Concepts — ACL 2021
Diego Antognini and Boi Faltings

Multi-Dimensional Explanation of Target Variables from Documents — AAAI 2021
Diego Antognini, Claudiu Musat, Boi Faltings

Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm — UAI 2021
Kirtan Padh, Diego Antognini, Emma L. Glaude , Boi Faltings , Claudiu Musat

Talks (selected)

Conversational Critiquing: From Recommender Systems to Text Generation — 2023
- Google Research, Zürich, Switzerland. — *Host: Dr. Claudiu Musat*

Efficient Machine Learning in Low-Resource and Highly-Specific Domains
- MIT-IBM Watson, Cambridge, MA, U.S.A. — *Host: Dr. Leonid Karlinsky*
- Swiss Text Analytics Conference 2023, Neuchâtel, Switzerland. — *Keynote*

Textual Explanations and Critiques in Recommendation Systems — 2022
- EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland. — *Host: Prof. Boi Faltings*

Interacting with Explanations through Critiquing — 2021
- University of Toronto, Online. — *Host: Prof. Scott Sanner*
- Swisscom AI, Lausanne, Switzerland. — *Host: Dr. Claudiu Musat*
- IJCAI 2021, Online.

Fast Critiquing with Self-Supervision for VAE-based Recommender Systems
- RecSys 2021, Online.

Rationalization through Concepts
- ACL 2021, Online.

T-RECS: a Recommender Generating Explanations and Integrating Critiquing — 2020
- ECAI 2020, Online.

Multi-Dimensional Explanation of Ratings from Reviews (Multi-Dimensional Rationalization)
- University of Zürich & NLP Meetup, Zürich, Switzerland. — *Host: Dr. Kornelia Papp*
- Swisscom AI, Lausanne, Switzerland. — *Host: Dr. Claudiu Musat*
- AAAI 2021, Online.

Learning to Create Sentence Semantic Relation Graphs for Multi-Document Summarization — 2019
- EMNLP 2019, Hong-Kong.

From Relation Extraction to Knowledge Graphs — 2017
- University of Applied Sciences, Neuchâtel, Switzerland. — *Host: Pr. Hatem Ghorbel*
- EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland. — *Host: Dr. J.-C. Chappelier*
- NLP Meetup, Zürich, Switzerland. — *Host: Dr. Kornelia Papp*

Honors & Awards

2023 **First plateau (i.e., 4 patents) invention achievement award**, IBM, Yorktown Heights, NY, U.S.A.
2023 **First patent application invention achievement award**, IBM, Yorktown Heights, NY, U.S.A.
2018 **First prize in the *IARPA Geopolitical Forecasting Challenge 2018***, macro-economics category, Washington, DC, U.S.A.
2014 **Excellent B.Sc. thesis award**, University of Applied Sciences, Neuchâtel, Switzerland.
2013 **Excelling B.Sc. student award**, University of Applied Sciences, Neuchâtel, Switzerland.

Interests

In my spare time, I ride motorbikes, dance salsa, drive boats, and paddle on beautiful Swiss lakes. I go to the gym regularly. I love reading and immersing myself in a wide range of subjects, such as leadership, communication, and finance. I have traveled to 30 countries and six continents.