

# Diego Antognini, PhD

## Research Scientist 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. Working on aligning large language models (LLM) with user preferences, building retrieval-augmented LLM systems for question answering on scientific corpora, and developing efficient models for resource-constrained training and inference. Experienced in designing explainable models that generate personalized and actionable textual explanations. Supervised 70+ B/M.Sc. projects.

## Skills

<b>Research Interests</b>	Generative AI, LLM alignment, efficient machine learning, self-supervision, question answering, recommendation.
<b>Program Committee</b>	NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, EACL, SIGIR, RecSys. <u>Journals</u> : ACL Rolling Review, ACM Computing.
<b>Languages &amp; Libraries</b>	<u>Efficient</u> : Python, PyTorch, Tensorflow, transformers, ONNX, Spark, Bash, SQL. <u>Prior Experience</u> : C++, CUDA, Java.
<b>Technologies</b>	GNU/Linux, Git, Poetry, Docker, Kubernetes, Openshift, API design, Redis, Elasticsearch, Milvus vector database.

## Experience

### Research Scientist

IBM Research AI

May 2022 - present

Zürich, Switzerland

- Publications: **5 papers in AI & ML leading venues**: 3 conference, 1 workshop, 1 under review at NeurIPS. Patents: **5 patents that are filed**.
- Designing methods to adapt and personalize LLMs to users, using parameter-efficient fine-tuning methods. To be integrated into *IBM Watsonx.ai*.
- Building low-latency models in the order of a few megabytes with similar performance and higher throughput than large models. Developed an unsupervised term extractor for technical domains and reduced the latency by 10x on a one-core CPU while performing similarly to BERT. Built an unsupervised term encoder using character-based models that match the quality of sentence encoders while being 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

Apr. 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).

### Machine Learning and Data Mining M.Sc. Thesis Intern

Iprova GmbH

Sep. 2016 - Mars. 2017

EPFL Innovation Park, Lausanne, Switzerland

- Developed a system to extract concepts from large corpora and build knowledge graphs for the invention team to gain new insights for patents.

## 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 8k+ 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)

<b>Assistive Recipe Editing through Critiquing</b>  Diego Antognini, Shuyang Li, Boi Faltings, Julian McAuley	EACL 2023
<b>pNLP-Mixer: an Efficient all-MLP Architecture for Language</b>  Francesco Fusco, Damian Pascual, Peter Staar, Diego Antognini	ACL 2023
<b>Extracting Text Representations for Terms and Phrases in Technical Domains</b>  Francesco Fusco* and Diego Antognini* (equal contribution)	ACL 2023
<b>Unsupervised Term Extraction for Highly Technical Domains</b>  Francesco Fusco, Peter Staar, Diego Antognini	EMNLP 2022
<b>Fast Critiquing with Self-Supervision for VAE-based Recommender Systems</b>  Diego Antognini and Boi Faltings	RecSys 2021
<b>Interacting with Explanations through Critiquing</b>  Diego Antognini, Claudiu Musat, Boi Faltings	IJCAI 2021
<b>Rationalization through Concepts</b>  Diego Antognini and Boi Faltings	ACL 2021
<b>Multi-Dimensional Explanation of Target Variables from Documents</b>  Diego Antognini, Claudiu Musat, Boi Faltings	AAAI 2021
<b>Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm</b>  Kirtan Padh, Diego Antognini, Emma L. Glaude, Boi Faltings, Claudiu Musat	UAI 2021

## Talks (selected)

<b>Conversational Critiquing: From Recommender Systems to Text Generation</b> • Google Research, Zürich, Switzerland.	2023 Host: Dr. Claudiu Musat
<b>Efficient Machine Learning in Low-Resource and Highly-Specific Domains</b> • MIT-IBM Watson, Cambridge, MA, U.S.A. • Swiss Text Analytics Conference 2023, Neuchâtel, Switzerland.	Host: Dr. Leonid Karlinsky Keynote
<b>Textual Explanations and Critiques in Recommendation Systems</b> • EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland.	2022 Host: Prof. Boi Faltings
<b>Interacting with Explanations through Critiquing</b> • University of Toronto, Online. • Swisscom AI, Lausanne, Switzerland. • IJCAI 2021, Online.	2021 Host: Prof. Scott Sanner Host: Dr. Claudiu Musat
<b>Fast Critiquing with Self-Supervision for VAE-based Recommender Systems</b> • RecSys 2021, Online.	
<b>Rationalization through Concepts</b> • ACL 2021, Online.	
<b>T-RECS: a Recommender Generating Explanations and Integrating Critiquing</b> • ECAI 2020, Online.	2020
<b>Multi-Dimensional Explanation of Ratings from Reviews (Multi-Dimensional Rationalization)</b> • University of Zürich & NLP Meetup, Zürich, Switzerland. • Swisscom AI, Lausanne, Switzerland. • AAAI 2021, Online.	Host: Dr. Kornelia Papp Host: Dr. Claudiu Musat
<b>Learning to Create Sentence Semantic Relation Graphs for Multi-Document Summarization</b> • EMNLP 2019, Hong-Kong.	2019
<b>From Relation Extraction to Knowledge Graphs</b> • University of Applied Sciences, Neuchâtel, Switzerland. • EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland. • NLP Meetup, Zürich, Switzerland.	2017 Host: Pr. Hatem Ghorbel Host: Dr. J.-C. Chappelier 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.