

# Diego Antognini, PhD

## Researcher in Artificial Intelligence & Machine Learning

✉ diegoantognini@gmail.com | 💻 www.diegoantognini.com | 🏠 Zürich, Switzerland | 🎓 Diego Antognini | 📊 Diego999 3.2k\* | 📄 diegoantognini

8 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

<b>Research Interests</b>	Generative AI, LLM alignment, multimodal, iterative refinement, efficient ML, NLP, conversational 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 Engineer

Jan. 2024 - present

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

### Research Scientist

May 2022 - Jan. 2024

- Publications: **10 papers in AI & ML leading venues**: 5 conference, 2 journal, 1 workshop, 2 demo. Patents: **5 filed patents**.
- Created data generation methods for aligning LLMs to convert multi-turn conversations into SQL queries for massive databases (*IBM FlowPilot*).
- Designed methods to adapt and personalize LLMs to users, using parameter-efficient fine-tuning methods. To be integrated into *IBM Watsonx.ai*.
- Built a distributed system to generate QA pairs using LLMs and a retrieval-augmented LLM to answer users' questions used in *IBM Deep Search*.
- 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.

### Module Head, Lecturer, M.Sc. Thesis Supervisor in NLP & LLMs

Feb. 2022 - present

- Designing and teaching two courses: deep learning for NLP and advanced generative AI (LLMs). Taught to 190+ unique M.Sc. students.
- Supervised 15 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

June 2015 - Dec. 2023

- Giving talks on a wide range of deep learning topics and offering machine learning consulting services for applied research in industrial projects.
- Assessed 40+ 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

Jul. 2021 - Nov. 2021

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

### Research and Teaching Assistant

May 2017 - Mar. 2022

- 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

Sep. 2017 - Mar. 2022

- 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

Sep. 2014 - Apr. 2017











- 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

Sep. 2011 - Aug. 2014

- 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>Trans-LoRA: towards data-free Transferable Parameter Efficient Finetuning</b> 	NeurIPS 2024
Runqian Wang, Soumya Ghosh, David Cox, <a href="#">Diego Antognini</a> , Aude Oliva, Rogerio Feris, Leonid Karlinsky	
<b>Paraphrase &amp; Solve: Exploiting the Impact of Surface Form on Mathematical Reasoning in LLMs</b> 	NAACL 2024
Yue Zhou, Yada Zhu, <a href="#">Diego Antognini</a> , Yoon Kim, Yang Zhang	
<b>MC Layer Normalization for calibrated uncertainty in Deep Learning</b> 	TMLR 2024
Thomas Frick, <a href="#">Diego Antognini</a> , Ioana Giurgiu, Benjamin F Grewe, Cristiano Malossi, Rong J.B. Zhu, Mattia Rigotti	
<b>Assistive Recipe Editing through Critiquing</b> 	EACL 2023
<a href="#">Diego Antognini</a> , Shuyang Li, Boi Faltings, Julian McAuley	
<b>pNLP-Mixer: an Efficient all-MLP Architecture for Language</b> 	ACL 2023
Francesco Fusco, Damian Pascual, Peter Staar, <a href="#">Diego Antognini</a>	
<b>Extracting Text Representations for Terms and Phrases in Technical Domains</b> 	ACL 2023
Francesco Fusco* and <a href="#">Diego Antognini</a> * (equal contribution)	
<b>Unsupervised Term Extraction for Highly Technical Domains</b> 	EMNLP 2022
Francesco Fusco, Peter Staar, <a href="#">Diego Antognini</a>	
<b>Fast Critiquing with Self-Supervision for VAE-based Recommender Systems</b> 	RecSys 2021
<a href="#">Diego Antognini</a> and Boi Faltings	
<b>Interacting with Explanations through Critiquing</b> 	IJCAI 2021
<a href="#">Diego Antognini</a> , Claudiu Musat, Boi Faltings	
<b>Rationalization through Concepts</b> 	ACL 2021
<a href="#">Diego Antognini</a> and Boi Faltings	

## Talks (selected)

<b>Conversational Critiquing: From Recommender Systems to Text Generation</b> <ul style="list-style-type: none"><li>Google DeepMind, Zürich, Switzerland.</li></ul>	2023 <i>Host: Dr. Claudiu Musat</i>
<b>Efficient Machine Learning in Low-Resource and Highly-Specific Domains</b> <ul style="list-style-type: none"><li>MIT-IBM Watson, Cambridge, MA, U.S.A.</li><li>Swiss Text Analytics Conference 2023, Neuchâtel, Switzerland.</li></ul>	<i>Host: Dr. Leonid Karlinsky</i> <i>Keynote</i>
<b>Textual Explanations and Critiques in Recommendation Systems</b> <ul style="list-style-type: none"><li>EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland.</li></ul>	2022 <i>Host: Prof. Boi Faltings</i>
<b>Interacting with Explanations through Critiquing</b> <ul style="list-style-type: none"><li>University of Toronto, Online.</li><li>Swisscom AI, Lausanne, Switzerland.</li><li>IJCAI 2021, Online.</li></ul>	2021 <i>Host: Prof. Scott Sanner</i> <i>Host: Dr. Claudiu Musat</i>
<b>Fast Critiquing with Self-Supervision for VAE-based Recommender Systems</b> <ul style="list-style-type: none"><li>RecSys 2021, Online.</li></ul>	
<b>Rationalization through Concepts</b> <ul style="list-style-type: none"><li>ACL 2021, Online.</li></ul>	
<b>T-RECS: a Recommender Generating Explanations and Integrating Critiquing</b> <ul style="list-style-type: none"><li>ECAI 2020, Online.</li></ul>	2020
<b>Multi-Dimensional Explanation of Ratings from Reviews (Multi-Dimensional Rationalization)</b> <ul style="list-style-type: none"><li>University of Zürich &amp; NLP Meetup, Zürich, Switzerland.</li><li>Swisscom AI, Lausanne, Switzerland.</li><li>AAAI 2021, Online.</li></ul>	<i>Host: Dr. Kornelia Papp</i> <i>Host: Dr. Claudiu Musat</i>
<b>Learning to Create Sentence Semantic Relation Graphs for Multi-Document Summarization</b> <ul style="list-style-type: none"><li>EMNLP 2019, Hong-Kong.</li></ul>	2019

## 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.