

Diego Antognini, PhD

Researcher in Artificial Intelligence & Machine Learning

✉ diegoantognini@gmail.com | 🌐 www.diegoantognini.com | 🏠 Zürich, Switzerland | 🎓 Diego Antognini | 📊 Diego999 3.3k+ | 📄 diegoantognini

9 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 90+ 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. <u>Journals</u> : ACL Rolling Review, ACM Computing.
Languages & Libraries	<u>Efficient</u> : Python, C++, PyTorch, Tensorflow, transformers, ONNX, Spark, Bash, SQL. <u>Prior Experience</u> : CUDA, Java.
Technologies	GNU/Linux, Git, Poetry, Docker, Kubernetes, Openshift, API design, Redis, Elasticsearch, Milvus vector database.











Experience

Google DeepMind <i>Zürich, Switzerland</i>	Jan. 2024 - present
Senior Research Engineer	Nov. 2024 - present
• Contributor of Gemini.	
Research Engineer	Jan. 2024 - Oct. 2024
• Contributor of Gemini.	
IBM Research, MIT-IBM Watson AI Lab <i>Zürich, Switzerland</i>	May 2022 - Jan. 2024
Research Scientist	
• <u>Publications</u> : 10 papers in AI & ML leading venues : 5 conference, 2 journal, 1 workshop, 2 demo. <u>Patents</u> : 5 filed patents .	
• Created data generation methods for aligning LLMs to convert multi-turn conversations into SQL queries for massive databases (<i>IBM FlowPilot</i>).	
• Designed methods to adapt and personalize LLMs to users, using parameter-efficient fine-tuning methods. To be integrated into <i>IBM Watsonx.ai</i> .	
• Built a distributed system to generate QA pairs using LLMs and a retrieval-augmented LLM to answer users' questions used in <i>IBM Deep Search</i> .	
• 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 <i>IBM Deep Search</i> to extract terms in real time from scientific documents and patents.	
Lucerne University of Applied Sciences <i>Lucerne, Switzerland</i>	Feb. 2022 - present
Module Head, Lecturer, M.Sc. Thesis Supervisor in NLP & LLMs	
• Designing and teaching two courses: advanced generative AI (LLMs) and deep learning for NLP. Taught to 190+ unique M.Sc. students.	
• Supervised 20+ M.Sc. theses in NLP with companies in the areas of medicine, law, politics, insurances, banks, media, and data visualization.	
HE-ARC – University of Applied Sciences <i>Neuchâtel, Switzerland</i>	June 2015 - Dec. 2023
Consultant and Expert for B.Sc. and M.Eng. Theses in ML	
• 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.	
UCSD – University of California San Diego <i>San Diego, CA, U.S.A.</i>	Jul. 2021 - Nov. 2021
Visiting Researcher in Prof. Julian McAuley's ML Lab	
• Published an unsupervised critiquing method for generative language models to help users rewrite cooking recipes to satisfy dietary restrictions.	
EPFL – Swiss Federal Institute of Technology in Lausanne <i>Lausanne, Switzerland</i>	May 2017 - Mar. 2022
Research and Teaching Assistant	
• 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	
EPFL – Swiss Federal Institute of Technology in Lausanne <i>Lausanne, Switzerland</i>	Sep. 2017 - Mar. 2022
Ph.D. in Computer Science	
• <u>Publications</u> : 15 papers in AI & ML leading venues : 8 conference, 6 workshop, 1 demo. <u>Advisor</u> : 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.	
• <u>Thesis</u> 📄: 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.	
EPFL – Swiss Federal Institute of Technology in Lausanne <i>Lausanne, Switzerland</i>	Sep. 2014 - Apr. 2017
M.Sc. in Computer Science	
• <u>Specialization</u> : 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.	
• <u>Thesis</u> : 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

- 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)

Trans-LoRA: towards data-free Transferable Parameter Efficient Finetuning 	NeurIPS 2024
Runqian Wang, Soumya Ghosh, David Cox, Diego Antognini , Aude Oliva, Rogerio Feris, Leonid Karlinsky	
Paraphrase & Solve: Exploiting the Impact of Surface Form on Mathematical Reasoning in LLMs 	NAACL 2024
Yue Zhou, Yada Zhu, Diego Antognini , Yoon Kim, Yang Zhang	
MC Layer Normalization for calibrated uncertainty in Deep Learning 	TMLR 2024
Thomas Frick, Diego Antognini , Ioana Giurgiu, Benjamin F Grewe, Cristiano Malossi, Rong J.B. Zhu, Mattia Rigotti	
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	

Talks (selected)

Conversational Critiquing: From Recommender Systems to Text Generation	2023
<ul style="list-style-type: none"> Google DeepMind, Zürich, Switzerland. 	Host: Dr. Claudiu Musat
Efficient Machine Learning in Low-Resource and Highly-Specific Domains	
<ul style="list-style-type: none"> MIT-IBM Watson, Cambridge, MA, U.S.A. Swiss Text Analytics Conference 2023, Neuchâtel, Switzerland. 	Host: Dr. Leonid Karlinsky Keynote
Textual Explanations and Critiques in Recommendation Systems	2022
<ul style="list-style-type: none"> EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland. 	Host: Prof. Boi Faltings
Interacting with Explanations through Critiquing	2021
<ul style="list-style-type: none"> University of Toronto, Online. Swisscom AI, Lausanne, Switzerland. IJCAI 2021, Online. 	Host: Prof. Scott Sanner Host: Dr. Claudiu Musat
Multi-Dimensional Explanation of Ratings from Reviews (Multi-Dimensional Rationalization)	2020
<ul style="list-style-type: none"> University of Zürich & NLP Meetup, Zürich, Switzerland. Swisscom AI, Lausanne, Switzerland. NLP Meetup, Zürich, Switzerland. AAAI 2021, Online. 	Host: Dr. Kornelia Papp Host: Dr. Claudiu Musat Host: Dr. Kornelia Papp
From Relation Extraction to Knowledge Graphs	2017
<ul style="list-style-type: none"> University of Applied Sciences, Neuchâtel, Switzerland. EPFL – Swiss Federal Institute of Technology in Lausanne, Switzerland. NLP Meetup, Zürich, Switzerland. 	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.