Julian Minder

STUDENT · COMPUTER SCIENTIST · DATA SCIENTIST

My research focuses on **NLP** and **interpretability**, particularly understanding the "black box" nature of AI systems. I aim to enhance model robustness and reduce bias, contributing to more trustworthy AI systems.

Education

ML Alignment and Theory Scholar BERKELEY (US) & LONDON (UK)
Jan. 2025 - August 2025 | Research Internship with **Neel Nanda**. Working on model diffing between the base and chat models.

MA Computer Science ETH ZÜRICH (CH)
Sept. 2021 - Dec. 2024 | major: Machine Intelligence, minor: Theoretical Computer Science

Master thesis: **"Mechanistic investigation of surfacing of hidden capabilities in Language Models"** – Supervisors: Dr. Chris Wendler (EPFL), Prof. Dr. Robert West (EPFL)
Conducted controlled experiments on toy models to investigate how fine-tuning reuses components from pretraining in language models. Building on these findings, a case study on large language models showed that fine-tuning can isolate a specific rank-1 subspace that governs the model's preference for parametric versus contextual knowledge. Notably, this subspace remained consistent across multiple versions of the same model. Our research provides evidence that fine-tuning primarily exploits this pre-existing subspace, providing concrete evidence for the reuse of existing hidden capabilities. OpenReview Full Thesis

BA Software Systems & Neuroinformatics UNIVERSITY OF ZÜRICH (CH)
Sept. 2017 - Sept. 2021 | major: Software Systems, minor: Neuroinformatics
Summa cum laude – Ø 5.7/6

Bachelor thesis: **"Enhanced String Similarity for Company Entities"** – Supervisors: Dr. Thomas Gschwind (IBM Research), Prof. Dr. Michael Böhlen (UZH) – Developed a method to improve the string similarity of company names by identifying information about semantic substructures in company names. Link to thesis. Grade: 6/6

International Exchange Semester UNIVERSITY OF UPPSALA (SE)
Jan. 2021 - July 2021

Communication Science & Media Research UNIVERSITY OF ZÜRICH (CH)
Sept. 2016 - July 2017 | major: Publicity & Communication, minor: Business Administration

Experience

ETH Zürich - Chair of Systems Design Student Research Assistant
Sept. 2021 (current position): 30-40%
Working on the DemocraSci project: Responsible for graph database design & data integration into knowledge graph (neo4j), developed python library for building data processing pipelines (data2neo, see projects). Primary contact person for the built parliament knowledge graph and for the data integration at chair. Independently explored various NLP approaches: Topic Modeling (BerTopic, fine-tuned T5 & SwissBERT models for text classification), experimenting with different training methodologies to create an embedding space for MPs to analyse ideology and other concepts.

IBM Research Zurich Research Intern
June 2019 - Dez. 2020: 30-100%
Record Linkage (C++/Python/Java) - Developed benchmarks and implemented company hierarchy matching. Developed API endpoints for services. Internship was extended from 6 months to 1 year due to interest on both sides. After Internship, I was approached by my team to write thesis at IBM Research Zurich.

AXA CC Robotics Lab Programmer
June 2018 - May 2019: 30-80%
Robotics Process Automation, Software Robots (C#), since Jan. 2019 responsible for the development of internal helper frameworks

Publications

Highlights

J. Minder*, K. Du*, N. Stoehr, G. Monea, C. Wendler, R. West, R. Cotterell (2024). Controllable Context Sensitivity and the Knob Behind It. Published as a conference paper at ICLR 2025. OpenReview.

J. Minder*, C. Dumas*, C. Juang, B. Chugtai, N. Nanda (2025). Robustly identifying concepts introduced during chat fine-tuning using crosscoders. ArXiv.

Conference Papers

J. Minder, F. Grötschla, J. Mathys, and R. Wattenhofer (2023). SALSA-CLRS: A Sparse and Scalable Benchmark for Algorithmic Reasoning. (Extended Abstract) Second Learning on Graphs Conference (LoG 2023).

T. Gschwind, C. Miksovic, **J. Minder**, K. Mirylenka, and P. Scotton (2019). Fast Record Linkage for Company Entities. 2019 IEEE International Conference on Big Data (Big Data), 623–630. IEEE.

In preparation

L. Brandenberger, S. Schlosser, L. Salamanca, L. Gasser, M. Balode, **J. Minder**, V. Jung, L. Babic, F. Perez-Cruz, and F. Schweitzer (2024). DemocraSci: A Knowledge Graph On the Swiss Parliament. Paper in preparation for Nature. Data Descriptor.

J. Minder, L. Brandenberger, L. Salamanca and F. Schweitzer (2024). Data2Neo - A Tool for Complex Neo4j Data Integration. arXiv.

S. Schlosser, **J. Minder**, and L. Brandenberger (2023). The Evolution of Parliamentary Polarization in Switzerland: An over century-long perspective. Paper Presentation at the ECPR Standing Group of Parliaments Conference, Vienna, AU, July 6-9, 2023., 1–11.

S. Schlosser, L. Brandenberger, **J. Minder**, G. Russo, L. Salamanca, and F. Schweitzer (2023). From Expertise to Versatility: The Evolution of Issue Engagement in the Swiss Parliament Over 130 Years. Paper Presentation at the European Political Science Association 13th Annual Conference, EPSA, Glasgow, UK, June 22-24, 2023.

D. Izzo, L. Salamanca, **J. Minder**, S. Schlosser, and L. Brandenberger (2023). How has Parliamentary Populism in Switzerland evolved over time? Evidence from the Swiss parliamentary speeches spanning 130 years. Paper Presentation at the European Political Science Association 13th Annual Conference, EPSA, Glasgow, UK, June 22-24, 2023.

J. Minder, L. Brandenberger, L. Salamanca, S. Schlosser, and A. Heidelberger (2023). Issue engagement in the Swiss parliament over the past 130 years. Paper Presentation at the at the Annual Convention of the Swiss Political Science Association, Basel, CH, February 2-3, 2023.

Research Projects

Semantic Uncertainty: Looking for meaning in embedding spaces
2023/ 2024

Exploration of methods for estimating semantic uncertainty of language models by fitting Gaussian Mixture Models onto samples of a language model. Small Semester Thesis working with Clara Meister and Niklas Stoehr at Prof. Dr. Ryan Cotterell's Lab. Grade 6/6.

SALSA-CLRS: An Algorithmic Learning Benchmark
2023

Development of a new benchmark for learning graph algorithms targeting sparse graphs and scalable architectures. Included analysis and training of various Graph Neural Network architectures. Semester Thesis at ETH at Distributed Computing Lab with Prof. Dr. Roger Wattenhofer. Passed (pass or fail). Result was published (see Publications). Thesis Report. Presentation Slides. LoG 2023 Poster. LoG 2023 Paper

Graph data integration library - Data2Neo
2022/ 2023

Development of a Python library for efficient integration of data from any source into a neo4j graph database. The library allows the user to define a dynamic data pipeline based on a simple yaml schema. In addition, the user can easily extend the basic functionality by adding their own pipeline Python functions. Data2Neo automates processing and automatically parallelizes on multiple cores. Written in connection with work at the the chair of system design. The library is used internally and is open sourced. Preprint.

Road Segmentation
2022

Presented an approach to use latent states of diffusion models to generate road segmentation masks for satellite images with very small training datasets. The diffusion model itself is trained on a large unlabelled dataset. Project for Computational Intelligence Lab at ETH (Spring 22). Tools: Pytorch, WandB. Grade 6/6.
Gitlab Repository (requires ETH Gitlab login). Report.

Human Pose Estimation
2022

Implemented and trained a model for predicting 3D human pose estimation from 2D images. Combined PARE Approach with ConvNeXt backbone. Project for Machine Perception course at ETH (Spring 22). Tools: Pytorch, WandB. Second place on class leader board of 30+ teams. Report.

DQN for Atari Pong
2021

Implemented and trained a DQN for Atari Pong with OpenAI gym. Project for Reinforcement Learning course at Uppsala University (Spring 21). Tools: Pytorch. Project Grade: 5/5. Github.

Musical Variation Automata
2019

Developed a Turing complete automata based on music notes and the concept of variation. Project for Models of Computation course at ETHZ (Spring 20). Implementation of compiler in python. Grade: 6/6. Github.

Full Semester Software Project - Santorini board game
2018

Server-Client system for Santorini board game. Responsible for system architecture and implementation of API and Backend. Tools: Java, Spring. Awarded for best project in FS19

Skills

Programming: Python, C++, C#, C, Java
Databases: SQL, Neo4j, BigQuery
Machine Learning and Data Science: Pytorch, PyG, Lightning, Transformers, Pandas, Numpy, WandB, NNsight, TransformerLens, Jax, Haiku
DevOps: Linux, Git, Docker, RUNAI

Languages

Native: German
Full Professional Proficiency: English
Limited Working Proficiency: French, Italian