Evžen Wybitul

Contact

E-mail: wybitul.evzen@gmail.com

GitHub: github.com/Eugleo

Education

ETH Zurich 2022 — (2025)

MSc in Data Science

GPA: 5.32 / 6.0 overall, 5.75 excluding pure math courses.

Courses include: Causality, Large Language Models, Natural Language Processing, Reliable and Trustworthy Artificial Intelligence, Data Science in Law and Policy.

Charles University 2021—2022

Auditing courses in MSc in Artificial Intelligence

Best possible grades in all courses I took.

Courses include: Reinforcement Learning, Evolutionary Algorithms, Probability Theory, Artificial Intelligence Theory.

Charles University 2018—2021

BSc in Bioinformatics

Top student in the programme in all three years, graduated with honors.

Courses include: Deep Learning, Mathematical Analysis, Linear Algebra, Data Structures.

Publications

Gradient Routing: Masking Gradients to Localize Computation in Neural Networks 2024

ArXiv, joint first author; MATS 6

A modification of back-propagation for learning specific capabilities in specific modules in the network, which can be used for unlearning and steering. Mentored by **Alex Turner** (Google DeepMind).

ViSTa Dataset: Do Vision-language Models Understand Sequential Tasks? 2024

ArXiV, first author; MATS 5

A dataset of 4,000+ videos of sequential tasks with descriptions. We use ViSTa to evaluate if visual-language models could serve as task supervisors in reinforcement learning. Mentored by **David Lindner** (Google DeepMind).

Refined SAEs: Transmuting Compute into Interpretability (2024)

Private draft available on request

An extension of sparse auto-encoders that uses test-time compute to produce better interpretable representations of the internal states of the model. Mentored by **Alex Turner** (Google DeepMind).

Other Research Experience

Assesing Vurneabilities in LLMs 2024

GitHub, course project under Florian Tramèr

Evaluated the safety of Large Language Model (LLM) agents, with a specific emphasis on prompt injections. Mentored by **Florian Tramèr**.

Training Steering Vectors 2024

PDF, course project

Produced first steering vectors for GPT-2 small. Explored the usage of sparse auto-encoder features for steering. Supervised by **Elliott Ash**.

Measuring Emotion in Political Language 2024

PDF, course project

Mapped how emotionality in political speeches developed over time. Supervised by **Elliott Ash**.

Certifying Robustness of Neural Networks 2023

GitHub, course project

Formulated an algorithm based on DeepPoly to certify neural network robustness against input perturbations.

Work Experience

ETH Zurich 2024

Teaching Assistant, Large Language Models

Taught a tutorial on the intuitions behind the transformer architecture and parameter-efficient finetuning methods.

IOCB Prague, research institute 2021

Assistant

1. Developed partial automation for a lengthy manual procedure that identifies cysteine bonds in proteins (thesis project).
2. Enhanced the reusability and accessibility of experimental data through the creation of an internal request management system.

MSD, pharmaceutical company 2019—2020

Junior data scientist

Contributed to cost reduction and increased drug yields by optimizing a complex drug preparation process.

Havířov Grammar School 2020—2022

Haskell curriculum developer & instructor

Developed and taught an introductory course in functional programming.

Software Projects

Technologies: Python (PyTorch), R, Julia, Haskell, Purescript, React, Typescript, PostgreSQL, Docker.

Hate Speech Detection in Online Comments

GitHub. Fine-tuned a BERT-based model for research and industry use in hate speech detection.

Request Management System for Researchers

GitHub. Provided a unified way to send, organize, and search through experiment requests and results.

Racket Language Extension for VS Code

GitHub. The most popular Racket extension for VS Code, with over 200 stars and 40,000 downloads.

Optimizing Exam Schedule

GitHub. A program designed to assist students in optimizing their exam preparation schedules.

Leadership Experience

1. In my project with David Lindner, I took the lead after the end of MATS to help push the project over the finish line.

   We had to change the research direction just before the end of MATS. I contributed a lot to the new direction of the project, aligned other colleagues to the idea, and often had individual calls with them to help them with next steps. I also wrote most of the paper, and created a majority of the videos.
2. I developed and taught a high-school functional programming course for two years. In some ways, this was highly similar to leading a project: motivating each individual person to do their piece of work, understanding their blockers, and finding ways to give them advice without invalidating their views.

Selected Awards and Achievements

Grant from Long Term Future Fund

Grant organization focused on beneficial research projects, often ones focusing on AI Safety.

Scholarship from Bakala Foundation

A highly selective scholarship, granted only to around five Master’s students in Czech Republic per year.

Scholarship for Outstanding Academic Achievement from Charles University

Awarded twice. I was the top student in my programme cohort, and top 6% student in the whole Faculty of Science.

Best A3 debater of the year

Finalist in the national team debating tournament, in which I was also recognized as the overall best A3 speaker.

References

- David Lindner, Google DeepMind
- Alex Turner, Google DeepMind
- Florian Tramèr, ETH Zurich