Evžen Wybitul

Contact

E-mail: wybitul.evzen@gmail.com

GitHub: github.com/Eugleo

Education

ETH Zurich Sep 2022 – (Aug 2025)

MSc in Data Science

GPA: 5.32 / 6.0 overall, 5.75 excluding pure math courses. Over 5.5 expected after Master's thesis.

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

Charles University Sep 2021 – Aug 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 Sep 2018 – Aug 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 backpropagation 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.

Other Research Experience

Assesing Vurneabilities in LLMs 2024

GitHub, course project

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.

Teaching Experience

ETH Zurich Feb 2024 – Aug 2024

Teaching Assistant, Large Language Models

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

Havířov Grammar School Sep 2020 – July 2022

Haskell curriculum developer & instructor

Developed and taught an introductory course in functional programming.

Work Experience

IOCB Prague June 2020 – July 2021

Assistant in a bioinformatics research group

1. Improved effectivity of a lengthy manual procedure that identifies cysteine bonds in proteins by partially automating it (thesis project).
2. GitHub. Built a web application for managing internal experiment requests (full stack web development). Enhanced the reusability and accessibility of experimental data.

MSD Sep 2019 – Feb 2020

Junior data scientist in a pharmaceutical company

Contributed to cost reduction and increased drug yields by optimizing a complex drug preparation process using classical ML on time-series data.

Selected 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.

Racket Language Extension for VS Code

GitHub. The most popular Racket extension for VS Code, with over 200 stars and 60,000 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.

   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 dataset.
2. I led a Bachelor's thesis in which we fine-tuned Llama for legal outcome prediction. The thesis received full marks from the committee.

Selected Awards and Achievements

Most Active Student Award, Bakala Foundation

Long-Term Future Fund Grant, EA Funds, $40 000

AI Safety Grant, AI Safety Support, $21 000

Scholarship, Bakala Foundation, $33 000

Scholarship for Outstanding Academic Achievement, Charles University, 2019

Scholarship for Outstanding Academic Achievement, Charles University, 2018

Team Finalist, National High-school Team Debating League

Best A3 speaker, National High-school Team Debating League

Absolute winner of a National Competition in Chamber Music

References

- David Lindner, Google DeepMind
- Ryan Cotterell, ETH Zurich
- Florian Tramèr, ETH Zurich