

Ilya Lasy

PreDoc Researcher


My research focuses on explainability and controllability of Large Language Models. In particular I'm interested in disentangling knowledge representations and achieving monosemanticity of hidden representations in LLMs by design. Currently I'm looking into approaching this problem from combining Mixture of Experts and Sparse Autoencoders.


 @ilyalasy

 @Ilya-Lasy

 @Misterion777

 @ilya.lasy

 ilya.lasy@tuwien.ac.at

 Vienna, Austria

 +43 660 7467971

Education

- PhD in Computer Science**
TUWien
09/2024
Informatics Faculty - Institute of Logic and Computation
Vienna, Austria
- MS in Computer Science**
Vilnius University
09/2020 – 06/2022
Faculty of Mathematics and Informatics
Vilnius, Lithuania
- BS in Software Engineering**
Belarusian State University of Informatics and Radioelectronics
09/2016 – 06/2020
Faculty of Computer Systems and Networks
Minsk, Belarus

Work experience

- Machine Learning Engineer (Part-time)**
Charisma.ai
10/2023
Building Large Language Models for cohesive interactive story generation, story quality evaluation, story structure extraction in creative entertainment industry.
Remote, London, UK
- Machine Learning Engineer**
wring.dev
04/2021 – 08/2023
Worked on ML-powered software testing automation product using Reinforcement Learning, Graph Neural Networks, Large Language Models.
Remote, Los Altos, California
- Machine Learning Engineer**
Adani Technologies
11/2020 – 02/2021
Developed Computer Vision solutions for healthcare and security using Python and frameworks (Pytorch, Tensorflow, OpenCV, etc.)
Minsk, Belarus

Publications

TU Wien at SemEval-2024 Task 6: Unifying model-agnostic and model-aware techniques for hallucination

In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), NAACL 2024

Dialogue System Augmented with Commonsense Knowledge

Lithuanian MSc Research in Informatics and ICT, 2022

Languages

English

C1

Belarusian

Native

Russian

Native

Achievements

EEML 2022 Attendee

Was selected to attend Eastern European Machine Learning Summer School 2022. Presented poster based on Master Thesis.

Master Thesis

Development of Open-domain Chatbot Augmented with Commonsense Knowledge

Bachelor Thesis

Implementation of text-to-speech system based on neural networks

SKILLS

Machine Learning

Solid knowledge of numerous classic ML/DL notions and algorithms.

Natural Language Processing

Text preprocessing, embeddings, language models, transformers, etc.

Reinforcement Learning

MDP, Policy/Value based methods

Python and ML/DL frameworks

Numpy, Pandas, Matplotlib, Scikit-learn, Pytorch, Tensorflow, ONNX