Ilya Lasy

PreDoc Researcher

My research focuses on mechanistic interpretability and controllability of LLMs.
In particular I'm interested in disentangling knowledge representations and achieving monosemanticity of hidden representations in LLMs by design. Currently exploring several (emergent) behaviours, such as memorization, reasoning and creativity, as case studies for understanding how knowledge is organized and accessed in LLMs.

@ilyalasy
@Ilya-Lasy
@Misterion777
@Misterion77

ilya.lasy@tuwien.ac.at | Vienna, Austria | +43 660 7467971

Education

PhD in Computer Science
TUWien
10/2023 — *Vienna, Austria*
Informatics Faculty - Institute of Logic and Computation

MS in Computer Science
Vilnius University
09/2020 – 06/2022 — *Vilnius, Lithuania*
Faculty of Mathematics and Informatics

BS in Software Engineering
Belarusian State University of Informatics and Radioelectronics
09/2016 – 06/2020 — *Minsk, Belarus*
Faculty of Computer Systems and Networks

Work experience

Machine Learning Engineer (Part-time)
Charisma.ai
10/2023 — *Remote, London, UK*
Building LLMs for cohesive interactive story generation, story quality evaluation, story structure extraction in creative entertainment industry.

Machine Learning Engineer
wring.dev
04/2021 – 08/2023 — *Remote, Los Altos, California*
Worked on ML-powered software testing automation product using Reinforcement Learning, Graph Neural Networks, Large Language Models.

Machine Learning Engineer
Adani Technologies
11/2020 – 02/2021 — *Minsk, Belarus*
Developed Computer Vision solutions for healthcare and security using Python and frameworks (Pytorch, Tensorflow, OpenCV, etc.)

Publications

Understanding Verbatim Memorization in LLMs Through Circuit Discovery
Proceedings of the First Workshop on Large Language Model Memorization (L2M2), ACL 2025

Guiding Generative Storytelling with Knowledge Graphs
Preprint: arXiv:2505.24803

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

Projects

LLM planning in stories
Work-in-progress on identifying whether LLMs plan ahead in creative tasks, such as story generation.

Hallucination Hackathon
Co-organized local hackathon to work on LLM hallucinations.

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 transformer based text-to-speech system

SKILLS

Mechanistic Interpretability
Circuit discovery, dictionary learning (SAE, CLT), attribution graphs
Practical: nnsight, transformer_lens

Deep Learning
Transformers, modern RNNs, MoEs, diffusion models, audio processing, embeddings, classic algorithms, etc.

LLMs
Post-training (SFT, RLHF, LoRa, etc)
Deployment (quantization, distillation, vLLM/sglang/llama.cpp)
Pipelines: RAG, vector DBs, haystack/llamaindex/langchain
Prompt Engineering: CoT, agents, dspy, textgrad, LLMlingua

Web Dev
Fullstack: streamlit, fastapi, flask, nodejs, vuejs, react
DevOps: docker, kubernetes, pulumi
Providers: AWS, Google Cloud