VLADIMIR IVANOV

EDUCATION

École Normale Supérieure Paris 2022-2024

- Bachelor's in Computer Science
- Bachelor's in Mathematics

Lycée Louis-le-Grand 2021-2022

- Undergraduate Mathematics & Computer Science & Physics
- (Classe Préparatoire MP2I)

PUBLICATIONS

Towards Safe Multilingual Frontier AI

- Study of the refusal rates of different models in the 24 official EU languages.
- My contribution was writing the code.
- arXiv GitHub

INTERNSHIPS

MATS (Redwood Research stream) Jan-Mar 2025

- Building realistic model organisms of alignment faking.
- Agentic reinforcement learning on frontier LLMs at scale.

ERA Fellowship affiliated with Cambridge University, Jul-Aug 2024

- Empirical study of deceptive and scheming tendencies in frontier LLMs.

SatisfIA Feb-Jun 2024

- Full-time technical AI safety internship.
- Practical and theoretical aspects of apsiration learning in deep reinforcement learning settings and known world model planning settings.

Realizability Interpretation of the Countable Axiom Of Choice Jun-Jul 2023

- Theoretical computer science research.
- Formalization in Coq.

TRAINING

Silver and Bronze International Mathematical Olympiad (IMO) Medals

- Also participated in other international mathematical olympiads (RMM, JBMO, MYMC).
- Participated in national computer science competition (Prologin).

ARENA Bootcamp Sep 2024

- One month long hands on empirical AI safety bootcamp.
- Interpretability, evals, reinforcement learning.
- PyTorch, TransformerLens, LLM APIs, Inspect AI.

ML4Good Camp by EffiSciences

· One week long hands on technical AI safety camp.
· PyTorch, LLMs, mechanistic interpretability, conceptual AI safety.

2022 Xena Project Undergraduate Workshop Imperial College, London

· One week long hands on Lean theorem prover workshop.
· My team worked on the Shannon-Lovász capacity theorem.

SELECT PROJECTS

Paper Replications

· Do Llamas Work in English? On the Latent Language of Multilingual Transformers. GitHub
· Towards Monosemanticity with Dictionary Learning. GitHub

Investigating the Influence of Dropout on Xor Probes in Neural Networks

· Small original research project, somewhat conclusive results.
· GitHub

Minuscule Stories

· Synthetic dataset of 320k stories much simpler than those in TinyStories.
· An 83k model trained on it speaks mostly coherent English.
· Github

PetitC Compiler

· Compiler of a subset of C.
· Semantics formalized and proven to be deterministic in Coq.
· Work in progress towards fully certifying the compiler.
· GitHub

EasyStenography

· Expands abbreviations in the keyboard input OS-wide.
· Github

Toy Self Replicating GPT4 Agent

· Toy scaffolding of GPT-4 with a terminal which managed to self replicate to a remote machine.
· Note that my definition of self replication is much weaker than what METR failed to elicit.
· GitHub

TEACHING

Olympiad math classes at Animath and Lyon Discrete Math Club.

Assistance with AGI Safety Fundamentals at École Normale Supérieure Paris

COMMUNITY ENGAGEMENT

Member of the AI Unit at EffiSciences

· EffiSciences is a nonprofit effective altruism organization funded by Open Philantropy.