Roman Kniazev

roman@knzv.me | https://roman.knzv.me
Paris, France

SUMMARY

AI researcher working on how neural networks learn, represent, and compute. Currently focused on mechanistic interpretability of transformers; previously on reinforcement learning, distributed systems, and logic. Comfortable across the stack, from theoretical analysis to designing end-to-end training pipelines in JAX and PyTorch.

EXPERIENCE

Post-doctoral researcher | *LaBRI, University of Bordeaux* Sep 2024 – Present

- **Conducted mechanistic interpretability research** on transformers trained on combinatorial reasoning tasks. Identified that the emergent world model mirrors the task's constraint algebra rather than its surface decomposition, and reverse-engineered the attention and MLP circuits implementing constraint propagation and forced placement.
- **Built BordAX** (github), a JAX-based modular deep RL framework with fully jitted training pipelines; 2.2–3.2x speed-up over StableBaselines3 with Gymnasium.
- **Engineered** a novel JAX implementation of gradient-based decision tree search, achieving **100x speed-up** over previous implementation.
- **Designed and delivered** a graduate course in *Deep Learning* for 30+ students.

Teaching and research assistant | *IRIF, Paris Cité University* Sep 2023 – Aug 2024

- **Created** materials and delivered undergraduate and graduate maths and CS courses.
- **Conducted** research on multi-player game semantics of distributed protocols.

PhD student | *LIX, Ecole Polytechnique* Sep 2020 – Aug 2023

- **Developed** new theoretical frameworks for concurrent and distributed computing using topological models to characterize emergent effects in multi-agent systems.
- **Taught** courses in concurrent and distributed computing, including multithreaded CPU programming and CUDA.

PUBLICATIONS

- **Transformers Linearly Represent Highly Structured World Models** (paper, code, blog)
  submitted to NeurIPS 2026 and MechInterp Workshop
- Computing the Reachability Value of Posterior-Deterministic POMDPs (2026) (arxiv)
- Semi-simplicial Set Models for Distributed Knowledge (2023) (arxiv)

Reviewer: JLC, JAIR, NeurIPS, POPL.

EDUCATION

Ecole Polytechnique | *PhD in Computer Science* 2023
Ecole Normale Superieure Paris-Saclay | *Master in Theoretical CS (MPRI), magna cum laude* 2020
Novosibirsk State University | *Bachelor in Mathematics and CS, summa cum laude* 2018

TECHNICAL SKILLS

- **ML stack:** JAX/Flax, PyTorch, HuggingFace Transformers.
- **Research areas:** Mechanistic interpretability, transformer architectures, reinforcement learning, post-training.