Roman **Kniazev**

https://roman.knzv.me
roman@knzv.me

SUMMARY

PhD in logic and distributed computing, now focused on AI and LLMs.
My background spans formal methods, RL, and distributed systems; my current work is mechanistic interpretability of LLMs and high-performance RL training in JAX.
I pick up new domains quickly and care about understanding what I build.
Seeking research/engineering roles in LLMs and AI.

EDUCATION

ECOLE POLYTECHNIQUE
PhD in Computer Science
2023 | Palaiseau, France

ENS PARIS-SACLAY
Master in Theoretical CS (MPRI)
2020 | Cachan, France
magna cum laude

NOVOSIBIRSK STATE UNIVERSITY
Bachelor in Maths and CS
2018 | Novosibirsk, Russia
summa cum laude

TECHNICAL SKILLS

PROGRAMMING
Proficient: Python, JAX, Java, Scikit-learn, Git
Familiar: PyTorch, HuggingFace Transformers, C++, Clojure, Rust

AI & ML
Reinforcement Learning, Deep Learning,
Transformers, RNNs, CNNs,
LLMs, Mechanistic Interpretability,
Formal Methods, Interpretable AI

SYSTEMS & THEORY
Concurrent & Distributed Computing Algorithms, Logic, Category Theory

MISC

Reviewer: JLC, JAIR,
NeurIPS 2025, POPL 2026

EXPERIENCE

POST-DOCTORAL RESEARCHER, AI & FORMAL METHODS
LaBRI, University of Bordeaux
Sep 2024 - now | Bordeaux, France
Working on *Safe AI through Formal methods* priority project with Nathanaël Fijalkow

- **Engineered** a novel implementation for gradient-based decision tree search in JAX, achieving a **100x** speed-up over previous methods for interpretable RL
- **Created** a JAX-based research-oriented modular RL training framework
- **Described** interpretable mechanisms in transformers solving combinatorial problems
- **Designed and delivered** a graduate course in *Deep Learning* for 30+ students
- **Co-authored** journal extension of a paper on decidable POMDP verification

TEACHING AND RESEARCH ASSISTANT
IRIF, Paris Cité University
Sep 2023 - Aug 2024 | Paris, France

- **Created materials and delivered** undergraduate and graduate maths and CS courses: programming, discrete mathematics, networks
- **Conducted** research on multi-player game semantics of distributed protocols

PHD STUDENT, LOGIC AND DISTRIBUTED COMPUTING
LIX, Ecole Polytechnique
Sep 2020 - Aug 2023 | Palaiseau, France
PhD thesis "On geometric models of epistemic logic"

- **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

SOFTWARE PROJECTS

RESIDUAL SUDOKU: [github] [blogpost coming soon]
A project in mechanistic interpretability investigating state representation and solving capabilities of causal transformers for sudoku puzzles

- Trained a GPT-2-style transformer on Sudoku solving traces (JAX, TPU)
- Found that linear probes on the residual stream recover both board state and per-cell candidate sets, confirming internal constraint propagation
- Found that representations peak at mid-layers and degrade in later layers, consistent with the model consuming its world model for token generation

BORDAX: [github]
Creator and lead developer of a JAX-based, high-performance, modular deep reinforcement learning framework, designed to support fully jitted training pipelines

- Supports highly customizable PPO and DQN, easy extendability
- 2.2x speed-up compared with StableBaselines3 with Gymnasium environments, 3.2x speed-up with jittable environments

SELECTED PUBLICATIONS

• Semi-simplicial Set Models for Distributed Knowledge
with Eric Goubault, Jérémy Ledent, Sergio Rajsbaum
Logic in Computer Science 2023 (rank A*)

• A many-sorted epistemic logic for chromatic hypergraphs
with Eric Goubault and Jérémy Ledent
Computer Science Logic 2024 (rank B)