Nikola **Jovanović**

Computer Science PhD Candidate · ETH Zurich

(+41) 77 987 92 47 | nikola.jovanovic@inf.ethz.ch | sri.inf.ethz.ch/people/nikola | LinkedIn | ni_jovanovic | Google Scholar

Research Interests

Safe & Trustworthy Machine Learning · Large Language Models

Latest Work

My current research focus is the problem of AI-generated text detection, primarily through the use of **LLM Watermarks**. This has so far led to one published work (ICML'24) and one preprint. Other publications are listed below, and an up-to-date list is maintained on Google Scholar.

project page — **Watermark Stealing in Large Language Models** — *ICML'24, Oral at R2-FM@ICLR'24*

Nikola Jovanović, Robin Staab, Martin Vechev

- We challenge claims regarding the robustness of LLM watermarks by showing that state-of-the-art schemes can be stolen for under $50, which for the first time enables realistic spoofing and scrubbing attacks at scale.

arXiv — **Black-Box Detection of Language Model Watermarks** — *under review*

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, Martin Vechev

- We show that a black-box adversary can rigorously test for the presence of an LLM watermark behind an API. Our tests reliably detect all 3 popular scheme families, despite efforts to mitigate such attempts by design.

Education

01.2022–present — **PhD Candidate, Computer Science**, SRI Lab, ETH Zurich, Expected graduation: 12.2025 — *Zurich, Switzerland*

- Advised by: Prof. Dr. Martin Vechev and Prof. Dr. Florian Tramèr.

2019 –2021 — **M.Sc. Computer Science**, ETH Zurich, Average grade: 5.85/6 — *Zurich, Switzerland*

2015 –2019 — **B.Sc. Computer Science**, Union University, Faculty of Computing, Average grade: 10.0/10.0 — *Belgrade, Serbia*

2011 –2015 — **Serbian Matura**, Mathematical Grammar School, Average grade: 5.0/5.0 — *Belgrade, Serbia*

Industry Experience

2019 (3mo) — **Software Engineer Intern, Camera Platform**, Snap Inc — *Los Angeles, USA*

- Implemented geometry understanding on point clouds to improve AR experiences. Prototyped a hybrid object tracking system based on latest research. Supported integration of neural face tracking algorithms.

2018 (3mo) — **Software Engineer Intern, Core Platform**, Improbable — *London, UK*

- Devised and implemented delta compression within SpatialOS, the main product of the company. Leveraged string algorithms and data structures to enable significant bandwidth savings on real-life client data.

2017 (3mo) — **Software Engineer Intern, Research & Machine Intelligence**, Google — *Mountain View, USA*

- Worked on semantic decomposition of Google Search queries in the setting of weakly supervised learning. Used an internal ML tool to fully automate training data curation and improve the quality of resulting data.

2016 (3mo) — **Software Engineer Intern, Knowledge Engine**, Google — *Zurich, Switzerland*

- Built an evaluation tool and a dashboard for Knowledge Graph entity deduplication algorithms. Manipulated large datasets using an internal parallel processing framework, an abstraction layer over MapReduce.

Honors and Awards

2024 — **Best Reviewer Award (Top 3%)**, ICML 2024 — *Vienna, Austria*

Oral Presentation (Top 3 Papers), ICLR 2024 Workshop on Reliable and Responsible Foundation Models — *Vienna, Austria*

- Awarded for the publication "Watermark Stealing in Large Language Models".

2023 — **Red Teaming Track Winner (1st Place, $60k Prize)**, US Privacy-Enhancing Technologies Prize Challenge — *Virtual*

- Awarded as part of the "ETH SRI" red team. The challenge was sponsored by NIST and NSF.

2021 — **Graduation with Distinction (MSc GPA above 5.75/6.0)**, ETH Zurich — *Zurich, Switzerland*

Honors, ACM ICPC World Finals 2019/2020, representing ETH Zurich — *Moscow, Russia*

Main Prize (2nd place, out of 1500+), IMC Trading 64BIDS Coding Challenge — *Virtual*

2020 — **Silver Medal (3rd place)**, ACM ICPC Southwestern Europe Regionals 2019/2020, representing ETH Zurich — *Paris, France*

Dositeja Scholarship for Postgraduate Studies Abroad, Young Talent Fund of Serbia — *Zurich, Switzerland*

2015 — **Full-Ride Scholarship**, Union University, Faculty of Computing, for success in Informatics competitions — *Belgrade, Serbia*

Bronze Medal, International Olympiad in Informatics (IOI) 2015 — *Almaty, Kazakhstan*

Bronze Medal, Balkan Olympiad in Informatics (BOI) 2015 — *Ruse, Bulgaria*

Supervised Students

MSc Student	**Thibaud Gloaguen**,	Black-Box Detection of Language Model Watermarks	*under review*
MSc Student	**Angéline Pouget**,	Back to the Drawing Board for Fair Representation Learning	*under review*
MSc Student	**Philipp Guldimann**,	Large-Scale LLM Benchmarking	
MSc Student	**Alexander Spiridonov**,	Large-Scale LLM Benchmarking	
MSc Student	**Robin Staab**,	From Principle to Practice: Vertical Data Minimization for Machine Learning, Training Data Extraction from Large Language Models	*IEEE S&P'24*
MSc Student	**Kostadin Garov**,	Hiding in Plain Sight: Disguising Data Stealing Attacks in Federated Learning	*ICLR'24*
Researcher	**Kamen Brestnichki**,	Gradient Leakage Attacks on GNNs in Federated Learning	
MSc Student	**Johannes Weidenfeller**,	Prompt Privacy in Large Language Models	

Teaching Experience

2023 –present **Rigorous Software Engineering (BSc)**, ETH Zurich 🔗, Exercise TA *Zurich, Switzerland*

2022 –present **Reliable and Trustworthy AI (MSc)**, ETH Zurich 🔗, Exercise and Head TA *Zurich, Switzerland*

- Exercise TA from 2022: Designing lectures, exercises, and exam questions, holding exercises.
- Additionally Head TA from 2023: Coordinating exercises, lectures and the exam, holding selected lectures.

2022 –2023 **Deep Learning for Big Code (MSc)**, ETH Zurich 🔗, Seminar TA (mentoring students) *Zurich, Switzerland*

2022 **Program Analysis for System Security and Reliability (MSc)**, ETH Zurich 🔗, Exercise TA *Zurich, Switzerland*

2021 **Eastern European Machine Learning Summer School**, EEML 🔗, Designed and held a Tutorial on GNNs *Virtual*

- An adapted version of the materials was open-sourced in Google DeepMind's educational repository 🔗.

2015 –2021 **Annual Alumni-led CS Seminar for Advanced High School Students**, MG Computer Science Week 🔗 *Belgrade, Serbia*

- Led the effort (2018-2021). Taught 6 lectures: Compilers and Functional Programming, Distributed Systems, Version Control, Dimensionality Reduction, Computational Geometry, Robustness of Neural Networks.

2019 **Machine Learning (BSc)**, Union University, Faculty of Computing 🔗, Exercise TA *Belgrade, Serbia*

- Created exercise materials for the course and open-sourced them on GitHub 🔗.

2018 **Computational Geometry (BSc)**, Union University, Faculty of Computing 🔗, Exercise TA *Belgrade, Serbia*

2017 **Object-Oriented Programming (BSc)**, Union University, Faculty of Computing 🔗, Student Helper *Belgrade, Serbia*

2016 **Introduction to Programming (BSc)**, Union University, Faculty of Computing 🔗, Student Helper *Belgrade, Serbia*

Publications

2024 **Black-Box Detection of Language Model Watermarks** *under review*
Thibaud Gloaguen, Nikola Jovanović, Robin Staab, Martin Vechev

Back to the Drawing Board for Fair Representation Learning *under review*
Angéline Pouget, Nikola Jovanović, Mark Vero, Robin Staab, Martin Vechev

Watermark Stealing in Large Language Models *ICML'24, Oral at R2-FM@ICLR'24*
Nikola Jovanović, Robin Staab, Martin Vechev

From Principle to Practice: Vertical Data Minimization for Machine Learning *IEEE S&P'24*
Robin Staab, Nikola Jovanović, Mislav Balunović, Martin Vechev

Hiding in Plain Sight: Disguising Data Stealing Attacks in Federated Learning *ICLR'24*
Kostadin Garov, Dimitar I. Dimitrov, Nikola Jovanović, Martin Vechev

2023 **FARE: Provably Fair Representation Learning with Practical Certificates** *ICML'23*
Nikola Jovanović, Mislav Balunović, Dimitar I. Dimitrov, Martin Vechev

2022 **LAMP: Extracting Text from Gradients with Language Model Priors** *NeurIPS'22*
Mislav Balunović*, Dimitar I. Dimitrov*, Nikola Jovanović, Martin Vechev

Private and Reliable Neural Network Inference *ACM CCS'22*
Nikola Jovanović, Marc Fischer, Samuel Steffen, Martin Vechev

On the Paradox of Certified Training *TMLR 10/2022*
Nikola Jovanović*, Mislav Balunović*, Maximilian Baader, Martin Vechev

Complete Verification via Multi-Neuron Relaxation Guided Branch-and-Bound *ICLR'22*
Claudio Ferrari, Mark Niklas Müller, Nikola Jovanović, Martin Vechev

2021 **Towards Robust Graph Contrastive Learning** *SSL@WWW'21*
Nikola Jovanović, Zhao Meng, Lukas Faber, Roger Wattenhofer

2018 **Towards Sparse Hierarchical Graph Classifiers** *R2L@NeurIPS'18*
Cătălina Cangea*, Petar Veličković*, Nikola Jovanović, Thomas Kipf, Pietro Liò