Nikolai Rozanov

UK resident, EU citizen
London, UK
nikolai.rozanov@gmail.com
https://nikolairozanov.com

SUMMARY

Experience: **Nikolai Rozanov** is an experienced **research scientist**, engineer and tech-entrepreneur. In his recent experience, Nikolai has been the **CTO and Head of Research** of an AI deep-tech start-up based out of London for 7 years. His experience spanned from deploying and productionising AI products, to state-of-the-art research and publishing papers in Tier-1 venues. He led research collaborations with leading research labs (including with **Prof. Iryna Gurevych** at TU Darmstadt, who was President of ACL during that time). Nikolai also led an *Innovate UK Grant* of £500K on an industrial AI project spanning a duration of 1.5 years. More recently, Nikolai is pursuing a full-time PhD at Imperial College London under the supervision of Dr Marek Rei on **LLM Agents & Reasoning**. Nikolai has also advised and consulted on various projects including a video gen-AI startup, Rask AI, where Nikolai led the LLM-based machine translation efforts as well as developed the machine translation evaluation suite for the company. Nikolai was also a speaker at various conferences and industry summits, such as CogX.

EDUCATION

2023-Present (Oct.-Present) **Imperial College London, London, UK**
PhD Candidate in the Computing Department

- LLMs, Foundation Models, Natural Language Processing (NLP), Reinforcement Learning (RL)
- Focus on Foundation Models and LLMs - AI-agents, reasoning, safety, hallucinations
- Research included a novel state-of-the-art algorithm that outperforms previous work by 15%
- Received professional training in CUDA programming (Oxford University & Imperial College)
- Awarded a full scholarship by the Department of Computing of Imperial College London

2017-2019 **University College London, London, UK**
MRes in Computer Science (Distinction)

- Awarded: Distinction (in the Computer Science Department)
- Thesis Topic was “Efficient Exploration in Deep Reinforcement Learning” (Distinction)
- Modules include: Statistical Learning Theory, Kernel Methods, DL & RL by Google Deepmind

2013-2016 **Imperial College London, London, UK**
B.Sc in Mathematics (First Class Honours)

- Awarded: First Class Honours, ARCS and received an academic award
- Attended a variety of classes to get an overview of Pure and Applied Mathematics
- Modules include: Topology, Analysis, Game Theory, Numerical Analysis and PDEs
- Also attended additional classes from the Department of Computing and a certified course in C

2005-2013 **Doblinger Gymnasium, Vienna, Austria**
Matura (Distinction)

- Graduated with the highest grades (Sehr Gut) across all subjects
- Focus lay within the Natural Sciences (Chemistry Olympiad, Physics Olympiad - Preparation)
- School President, notably elected with a 75% majority of votes

PROFESSIONAL EXPERIENCE

2025-Present (Aug.-Present.) **MBZUAI, Abu Dhabi, UAE**
Visiting Student supervised by Prof. Iryna Gurevych and Prof. Preslav Nakov

- Working on LLM coding agents for the HPC Benchmark “KernelBench”
- Improving well-known and widely used AIDE algorithm for GPU programming
- Won prestigious Weco AI Fellowship for pioneering work on automatic GPU programming
- Actively supervised several LLM Agent projects

Nikolai Rozanov

UK resident, EU citizen
London, UK
nikolai.rozanov@gmail.com
https://nikolairozanov.com

2023-Present (Oct.-Present.) **Imperial College London, London, UK**
Graduate Teaching Assistant & PhD Rep

- Taught several Master's level modules: Reinforcement Learning, Computer Vision, NLP
- Organised the departmental conference "Imperial Computing Conference" (ICC24S, ICC24W)
- Liaise with the director of research as head rep to improve the post-graduate experience
- Received professional training in teaching (pursuing Fellowship)
- Additional coach for the International Collegiate Programming Competition (ICPC) Team

2024 (Feb.-Dec.) **Rask AI Inc., London, UK**
Senior Research Scientist (Contractor, Part-time)

- Joined a dynamic video genAI start-up with 2M+ users and 5M+ ARR as senior scientist
- Developed the evaluation framework for the entire company for machine translation (MT)
- Published a resource and metrics paper to WMT24 (EMNLP2024) on iso-chronic translation
- Co-developed the LLM-based translation engine; and audio ML solutions (diarization, ASR)
- Represented Rask AI at various executive events, including "The Podcast Show" & EMNLP

2017-2023 (Feb.-Oct.) **Wluper Ltd., London, UK**
Co-Founder, CTO, Research Scientist

- Leading the tech team of 9 senior ML/SWE Engineers and Scientists (ml, FE/BE, devops)
- Developing core system: algorithms & backend (Python + ML), frontend (JS) and devops (aws)
- Leading collaborations with major research labs and our senior research team (PhDs, Post Docs)
- Co-published six (6) papers during this time, some in collaboration with major research labs
- Co-leading product definition and requirements based on client and market needs
- Awarded Innovate UK Research Grant 2019 (£500,000.00)

2016-2017 (Sep.-Feb.) **Rezonence, London, UK**
Software Developer (Part-time)

- Worked on automating tools for detecting advertisements on websites
- Focused on NodeJS and Javascript development
- Learned about Amazon Cloud computing and software development in a team

2016 (Jun.-Sep.) **Imperial College London, London, UK**
Summer Research Intern in Machine Vision

- Worked on Aruco marker tracking using OpenCV and the Aruco library
- Developed and extended C++ skills for creating reusable and platform independent code
- Worked on combining Computer Vision, Projective Geometry with existing Aruco library

2015 (Jun.-Aug.) **Imperial College London, London, UK**
Summer Research Intern in Monte Carlo Methods

- Optimised the Metropolis Hastings Algorithm using variance reduction techniques
- Developed very efficient implementation of Antithetic MCMC for Stochastic PDES

2012 (Aug.-Sep.) **Bank Gutmann AG, Vienna, Austria**
Summer Banking Intern in Private Equity

- Analysed market opportunities in the private equity department
- Presented findings in a 1-1 to the CEO of the bank

2011 (Aug.-Sep.) **Attesta AG, Vienna, Austria**
Summer Accounting Intern

- Learned the basics of accounting, financial reporting, tax reporting and audits
- Helped in conducting an audit of one of Austria's major political parties

PUBLICATIONS

Google Scholar: https://scholar.google.com/citations?user=fi-feOEAAAAJ&hl=en

Nikolai Rozanov

UK resident, EU citizen
London, UK
nikolai.rozanov@gmail.com
https://nikolairozanov.com

2025 (Published) **StateAct: Enhancing LLM Base Agents via Self-prompting and State-tracking**
2025 Workshop for Research on Agent Language Models (REALM 2025 @ ACL 2025)
- "StateAct: Enhancing LLM Base Agents via Self-prompting and State-tracking"
- We managed to create our own version of "chain-of-thought", called "chain-of-states"
- Our method achieves +15% against the previous best method ReAct
- On par performance with RAG & code-execution without using RAG or code
- Method allows to check for self-consistency type hallucinations within LLM generation

2024 (Published) **IsoChronoMeter: A simple and effective isochronic translation metric**
2024 Conference on Machine Translation (WMT 2024 @ EMNLP 2024)
- "IsoChronoMeter: A simple and effective isochronic translation evaluation metric"
- Developed a first automatic metric to measure iso-chrony in (automatic) translation
- Created a high-quality evaluation suite for WMT24 for testing iso-chrony for MT

2023 (Published) **Learning From Free-Text Human Feedback**
2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
- "Learning From Free-Text Human Feedback--Collect New Datasets Or Extend Existing Ones?"
- A taxonomy of user-led corrections in dialogue systems and a method to find them in datasets

2022 (Published) **Self-alignment Training for Auto Encoders**
21st International Conference on Machine Learning Applications (ICMLA 2022)
- "Connecting the Semantic Dots: Zero-shot Learning with Self-Aligning Autoencoders and a New Contrastive-Loss for Negative Sampling"
- Novel simple zero-shot learning method
- Achieved State-of-the-art (SOTA) results & code is open source on Github

2021 (Published) **MATILDA - Dialogue Annotation Tool**
16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)
- "MATILDA: Multi-AnnoTator multi-language Interactive Light-weight Dialogue Annotator"
- Dockerised annotation tool with persistent MongoDB database

2020 (Published) **Adapter-Based Knowledge Infusion into Transformer Language Models**
2020 Conference on Empirical Methods in Natural Language Processing (DeeLIO 2020, EMNLP 2020)
- "Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers"
- Novel method of transferring knowledge bases into transformer based language models
- Achieved big gains (~20 performance points over BERT) in Diagnostic GLUE tasks

2019 (Published) **LIDA - Dialogue Annotation Tool**
2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)
- "LIDA: Lightweight Interactive Dialogue Annotator"
- First purpose-built annotation tool for research dialogue annotation (single researchers)

2018 (Published) **Evolutionary Data Measures**
22nd Conference on Computational Natural Language Learning (CoNLL 2018)
- "Evolutionary Data Measures: Understanding the Difficulty of Text Classification Tasks"
- Evolutionary algorithm to evolve from 50 base metrics a novel composite metric
- Achieved high correlation on task difficulty and model performance (~0.85 Pearson Corr.)

PERSONAL AWARDS (selected)

2025 (6 months) **Weco AI Fellowship** - Fellowship to further develop the AIDE algorithm on KernelBench
2025 (6 months) **Anthropic Research Grant** - Research Grant from Anthropic to develop LLM Agents
2024 (6 months) **OpenAI Research Grant** - Early Researcher Grant from OpenAI to develop LLM Agents
2024 **NeurIPS Reviewer Award** - Free attendance of NeurIPS Conference
2022 **ICML Reviewer Award** - Free attendance of ICML Conference

Nikolai Rozanov

UK resident, EU citizen
London, UK
nikolai.rozanov@gmail.com
https://nikolairozanov.com

ADDITIONAL PROJECTS

2024-Present **Accelerated (GPU) Programming (CUDA, HIP, C/C++)**
- Oxford University Course on CUDA & HPC with Prof. Mike Giles;
- NVIDIA & Imperial College London Course on multi-gpu CUDA programming (certified)
- Low-level CUDA Kernels (Bandwidth vs. Compute limited); GPU/HPC architectures;

2019-Present **Custom-built computer & Linux/Unix Administration**
- (Partially) custom-built PC (multiple-GPUs) and Linux administration (turning into remote VM)
- Learned the basics of hardware and Unix-based operating systems; doing sys-admin, patching,...

2014-2022 **Reinforcement Learning & Probabilistic Machine Learning (Python)**
- Created baselines and evaluation suites for "exploration" for reinforcement learning (RL)
- Methods involved: Kernel methods, monte-carlo methods, message passing algorithms,...
- Implemented various machine learning algorithms: EM, Kalman-filters, BP, HMMs, MCMC,...

2016 **Computer Vision (C++, OpenCV, OpenGL)**
- Built a 3D simulator in OpenGL for tracking Aruco markers
- Implemented portable framework that interfaces between OpenGL and OpenCV

2015-2016 **High Performance Computing & Scientific Computing (Python, C, C++, Matlab)**
- Imperial College London courses focusing on OpenMPI, OpenMP and scientific computing
- Learned a variety of methods in parallel distributed systems for complex computations
- Developed a high-precision numerical library in low-level C for solving polynomials, FFT, etc.
- Implemented "Finite Difference Schemes" for dynamical systems on non-convex domains

2013-2015 **Other (C, arm-Assembly, Haskell)**
- Developed an assembler for arm-arch in C and built simple programs in Assembly
- Worked on Project Euler using Haskell, implemented concise algorithms for Primes and data.

PROGRAMMING SKILLS (Summary)

2009-Present **Programming Languages**
- **Python** (10+ years): Many ML & SWE projects (torch, transformers, flask, unsloth, vllm...)
- **Javascript, NodeJS, VueJS** (3+ years): Various frontend and backend projects (pinia, wss)
- **Linux/Unix/Bash** (10+ years): General sys-admin, bash scripts, package management
- **Docker, DevOps** (1-2+ years): AWS - DynamoDB, IaaC, CI/CD, Lambda, Gateway, S3, EC2
- **C/C++** (4+ years): particular focus on improving speed, numerical precision, parallelism
- **Matlab** (3+ years): A wide variety and depth of projects (modeling, prediction etc.)

2013-Present **Libraries and Frameworks**
- **Pytorch, Tensorflow** (8+ years): Convs, RNNs, transformers etc. to solve various ML tasks
- **CUDA, Triton** (0-1 years): Custom low-level kernels (shuffles, shared-mem); mem vs. compute
- **Transformers, Unsloth, Sklearn, Numpy, Pandas** (1-3+ years): Various ML, NLP projects
- **OpenAI, Anthropic, Cohere, Huggingface** (3+ years): Various LLM projects
- **OpenGL, OpenCV** (0-1 years): Graphics shaders, marker detection, camera transformation
- **OpenMP, OpenMPI** (0-1 years): Focused on creating high-speed computational models

EXTRA CURRICULAR ACTIVITIES

2024 "**The Podcast Show 2024" Speaker** Spoke about automatic AI-based dubbing
2023 "**CogX 2023" Speaker** Spoke about LLMs and ML benefits to industrial use-cases
2022 **PlugAndPlay Accelerator Speaker (London)** Spoke about the benefits of Voice AI in industry
2019 "**Forbes 30 under 30" Award** in the category European Technology Section (with Wluper Ltd.)
2012 **Olympiads** Chemistry Olympiad (3rd prize Vienna), Philosophy Olympiad (3rd place Vienna)
2012 **Oratory** Finalist at a nationwide speech competition (Economics and Integration Association)
2002-Present **Chess** Medalist in many international, national and regional competitions (e.g. Gold BUCA Plate)
1995-Present **Languages: German** (native), **English** (fluent), **Russian** (fluent speaker), **French** (basic), **Latin** (basic)
2005-Present **Sports: Swimming**, **Ice Hockey**, **Sailing**