

Vadim Fomin

hi@wadimiusz.me

[Github](#) | [LinkedIn](#)

[Google Scholar](#) | [Substack](#)

Summary

ML engineer pivoting to AI safety research. I expect misalignment by default and think only a thorough understanding of a system can justify confidence in its alignment, so I focus on empirical work, especially interpretability.

Doing NLP continuously since undergrad (2016). Experience with 0-to-1 projects in both research and industry. Master's thesis (2020) probed how discourse structure is represented across transformer layers.

Projects

[MoSSAIC \(AISC 2026\)](#)

Searching for interpretability techniques with genetic programming, aimed at porting across architectures.

[Theorem marketplace](#)

Built and deployed a dApp (Ethereum Sepolia testnet) where one can place a bounty on a formal proof of a theorem, claimable by submitting the proof. Solo, 0-to-1.

Publications

7 peer-reviewed papers, 2 as first author; 85 citations, h-index 4 (Google Scholar).

First author: Dialogue 2019 (cultural semantic shifts in Russian, arXiv:1905.06837); Dialogue 2018 (spelling correction of GIS search queries).

Co-author: LREC 2020 (AI Journey overview); Dialogue 2020 (Shiftry); LChange workshop at ACL 2019 (evaluative adjectives); AIST 2018 (text augmentation); FigLang workshop at NAACL 2018 (metaphor detection).

Work Experience

Independent AI safety research (2026 — current)

Aiming to extract formal-language descriptions (automata) directly from transformer weights, with Tracr-compiled models as ground truth. Pilot write-up: wadimiusz.me/nerode; code: github.com/wadimiusz/nanogpt.

Volta Health (2025 — 2026)

- medical scribe for patient visits
- dialogue agent for compliance navigation with internal MCP retrieval of user-specific information
- OCR and translation pipeline for immunization documents in user's original language

ModelFront — ML engineer (2022 — 2025)

Solo ML engineer since 2023. Owned the infrastructure of processing the data, training models for all our products, and serving them.

Created many data augmentation techniques for making the models robust to inputs not seen in training data, using NLP controllability methods among others, nearly doubling the number of customers. Cut the inference time of encoder-decoder models by around 3 times. Worked closely with the engineering team to keep ML and production close.

[Computational Pragmatics Lab](#) — **Research Assistant (2020)**

Probing transformer layers for discourse structure.

Sber — NLP Engineer (2019 — 2020)

- Developed dialogue systems.
- Helped organize and develop a baseline for a national competition in natural language understanding and generation.

Education

[Technical AI Safety](#), BlueDot Impact (2026)

Université de Lorraine / LORIA, Nancy, France (2020 — 2022)

Ph.D. candidate (left to join a startup)

Research on multi-factor controllability of generation for fictional dialogs (personality, relationship, setting, emotion etc.). Research and implementation of different techniques of steering generated texts into necessary directions and evaluating outputs.

Computational Linguistics, HSE (2018 — 2020)

Master's thesis: linear probes for discourse structure across transformer layers, two languages, three models.

Mathematical Linguistics, Novosibirsk State University (2014 — 2018)

BA thesis in NLP (spell-checking search queries). Active in the local NLP group since its birth. Published 3 of the papers listed under Publications, participated in hackathons and took a prize in one.