Dmitrii Kuzmin

NLP/ML Engineer · NLP Researcher

1kkiren@mail.ru | 1kkiren.ru | 1kkiRen

Summary

NLP/ML Engineer with experience in both research and product work on LLMs. Fine-tuning (SFT, LoRA/QLoRA) and inference of open-weight models (Qwen2.5-VL, Kimi2.6), development of agentic pipelines (LangGraph, LangChain, LiteLLM), research in tokenization and LLM uncertainty estimation. Publications: AI Journey 2025 (accepted), 2 papers at A* conferences (under review). Experience with GPU clusters (NVIDIA, MetaX), CUDA, Docker, Git.

Skills

LLM / NLP	PyTorch, Hugging Face Transformers, PEFT (LoRA/QLoRA), SFT, tokenization (BPE, SentencePiece), vLLM, Triton
Agentic & RAG	LangChain, LangGraph, LiteLLM, Pydantic, Qdrant
MLOps / Infra	CUDA, Docker, Git, Linux, Bash, W&B, distributed inference on GPU (NVIDIA, MetaX)
Data & Backend	Python, NumPy, Pandas, scikit-learn, SQL, FastAPI
Languages	English (C1, confirmed), Russian (native)
Soft skills	Flexibility, Responsibility, Enthusiasm

Work Experience

Neural Systems and Deep Learning Lab, MIPT

Moscow, Russia

NLP Engineer — *May 2025 - Present*

- Maintaining and scaling LLM inference on 3 clusters of 17 GPUs (NVIDIA A100/H100 + MetaX), serving 6 models (from 14B Dense to 1T MoE) with 150k+ req/day load; p95 latency reduced from 3000 ms to 2500 ms (17%).
- Leading research on uncertainty estimation in the reasoning of large language models.
- Designed and deployed an agentic pipeline (LangGraph) for automated generation of Selenium tests from natural language descriptions.
- Designed the architecture of a co-pilot solution for automating call center operators' work.
- Designed an LLM uncertainty evaluation benchmark (40+ models, 4 domains).
- Launched a regular LLM-eval (LLM-as-a-Judge, benchmark evaluation criteria) pipeline on MetaX GPUs.

Mohamed bin Zayed University of Artificial Intelligence

Abu Dhabi, UAE (remote)

Intern Researcher, NLP — *June 2025 - Present*

- Conducting research on adapting LLM tokenizers to Russian; experiments on Qwen2.5-1.5B-Instruct, 1000+ GPU-hours, achieving a 17% reduction in fertility relative to the baseline tokenizer.
- Coordinating experiments with research teams.
- Researched an impact on glitch tokens on model generation.

Center for Applied AI, Skolkovo

Moscow, Russia

Middle NLP Engineer — *February 2025 - May 2025*

- Fine-tuned Qwen2.5-VL (LoRA, r=16, 3 epochs) on a corpus of 200 building drawings for object detection: mAP@0.5 = 0.6, mAP@[.5,.95] = 0.35.
- Assembled an e2e pipeline for object detection in drawings (preprocess → inference → postprocess → report) with 87% class coverage.
- Selected and A/B tested prompts for generating remarks on drawings.

Higher School of Economics

Moscow, Russia

Research Assistant, NLP — *June 2024 - May 2025*

- Fine-tuned Llama3-8B-Instruct (SFT, 300 GPU-hours); achieved an 11.6% improvement on the overall MERA benchmark score.
- Developed a BPE tokenizer for Russian on an 80 GB corpus; reduced the average number of fragments per word from 3.86 to 1.5 (a 61% reduction).
- Implemented methods for manipulating existing tokenizers (token deletion/addition/merging) — the open-source library TokenizerChanger (PyPI).
- Created a grammar benchmark for Russian, used in the AI Journey 2025 publication.

Moscow Aviation Institute

Moscow, Russia

ML / Backend Engineer — *July 2023 - October 2023*

- Developed a sentence topic classification model with 10+ classes.
- Optimized SQL queries and database schema: p95 latency reduced from 1500 ms to 800 ms (a 47% reduction).

Innopolis University

Innopolis, Russia

NLP Engineer (Internship) — *June 2023 - July 2023*

- Fine-tuned RuBERT for sentiment analysis of YouTube comments on a dataset of 10000 examples; F1 = 0.82.

Publications

A Multi-Aspect Evaluation of Tokenizer Adaptation Methods for LLM on Russian

AI Journey 2025 — accepted (poster)

Co-author *2024–2025*

- Systematic comparison of 4 BPE tokenizer adaptation methods for Russian-language LLMs on 3 benchmarks (MERA, MMLU, TA bench) and additional statistical methods.
- The best method reduces fragmentation from 3.7 to 2.34 (-36.7%).
- `https://www.mathnet.ru/eng/danma690`

Mitigating the Impact of Glitch Tokens via Targeted Retokenization

AACL 2026 — under review

Co-author *2025–2026*

- An inference-time retokenization method that removes under-trained (“glitch”) tokens based on embedding norm/entropy; does not require model fine-tuning.
- Evaluated on 6 models (from 1.5B to 32B) and the MMLU and WMT benchmarks.

TokenSubstitution: Cost-Efficient Method of Language Adaptation Based on Token “Trained-ness”

EMNLP 2026 — under review

First author *2025–2026*

- A method for adapting an LLM to a new language by replacing weakly trained tokens with tokens of the target language, without full embedding fine-tuning.
- -15% inference latency while surpassing the original model on the MERA benchmark.

Projects

Tokenizer Changer

Library *Jul. 2024 - present*

- The open-source library for changing existing tokenizers.
- `https://pypi.org/project/TokenizerChanger/`
- `https://github.com/1kkiRen/Tokenizer-Changer`

CRUD Calendar LLM Chatbot

Freelance *Feb. 2025*

- Telegram-chatbot that works as a calendar manager and can provide an up-to-date news summary.

Education

Innopolis University

Innopolis, Russia

B.S. in Data Analysis and Artificial Intelligence *2022 - 2026*

- Core Courses: Software Systems Analysis and Design, Human AI Interaction Design, Mathematical Analysis

Extracurricular Activity

Innopolis University

Innopolis, Russia

Tutor *Sep. 2023 - Jan. 2024*

- Helped 1st Bachelor’s students adapt to the University.
- Organized extracurricular events.