

Dmitry Balobin

Senior NLP / LLM Engineer

Open to opportunities in the UK and Serbia
Available for relocation

hanaan@internet.ru
linkedin.com/in/d0rj
huggingface.co/d0rj

Summary

Senior NLP / LLM Engineer with experience building, fine-tuning and deploying large language model systems across search, banking, maps and conversational AI products. Strong background in LLM training pipelines, RLHF, SFT, RAG, summarisation, AI agents and production ML infrastructure. Experienced in research experimentation, applied NLP product development and technical interviewing.

Work Experience

Yandex

Jan 2026 – Present

Senior ML Engineer, Alice on Search

- Optimised LLM training infrastructure on 32–64 GPUs, making training more than 2× faster (huggingface: yandex/AliceAI-T5-35B-A0.6B).
- Optimised the production model, increasing throughput (RPS) by 5.6%.
- Reduced the runtime of the model optimisation pipeline before production deployment by approximately 40%.
- Researching encoder-decoder and decoder-only MoE LLM architectures.

Central University

Jun 2026 – Present

Mentor

- Supporting students' academic and career development.

Raiffeisen Bank

Sep 2025 – Jan 2026

Senior AI Engineer, Corporate & Investment Banking

- Built two projects from scratch to pilots in four months — AI Copilot and autonomous agents for online trading.

2GIS

Jan 2024 – Sep 2025

Senior NLP Data Scientist

- Independently developed and deployed review summarisation (the “People Say” feature), increasing click-through conversion.
- Worked on developing generative summarisation for online search.
- Developed product-to-tag matching with macro F1 = 0.97 on a human-annotated dataset.
- Conducted technical interviews across NLP and adjacent ML teams, assessing candidates in NLP, ML engineering and applied data science.

Sber

Jun 2022 – Dec 2023

ML Engineer, NLP B2C Team

- Developed the RLHF training stage for the GigaChat LLM and ran research experiments to improve model behaviour and alignment.
- Improved the SFT training pipeline for GigaChat and adapted internal models to product-specific requirements.
- Built pilot integrations of GigaChat into B2C and internal banking products.
- Developed retrieval components for RAG-based business use cases powered by GigaChat.

Sber

Dec 2021 – May 2022

Data Scientist Intern, NLP B2C Team

- Worked on speech analytics for telephone calls, including clustering and visualisation of call data.
- Improved the intent recognition model training pipeline and adapted it to internal quality requirements.

Education

MIPT, School of Applied Mathematics and Informatics

2025 – 2028

*PhD Candidate, Computer Science and Informatics***MIPT, School of Applied Mathematics and Informatics**

2023 – 2025

*MSc, Computer Science; Machine Learning and Data Analysis***MIREA – Russian Technological University**

2019 – 2023

*BSc, Applied Mathematics; Data Analysis***Publication**

D. Y. Balobin, O. D. Somov. Improved Detection Methods for Generated Texts

2025

Technical Skills

Frameworks & Tools: PyTorch, Transformers, DeepSpeed, verl, vllm, LangChain, PydanticAI, Langfuse, Kubernetes, FastAPI.