Belov Vitalii

Moscow, Russia | +7 967 961 58 24 | vitalii.belof@gmail.com | GitHub: poliroika | Telegram: @poliroika

Target: AI/ML Engineer (Conversational AI). Open to relocation to Japan; visa sponsorship required.

Summary

AI/ML engineer (3+ yrs) focused on **LLMs/RAG** and prod NLP. Built e2e retrieval → prompt templates → extraction/validation → evaluation. Designed **Graph RAG** (meta-chunk linking; section proximity/co-occurrence) and integrated vector search (FAISS/Qdrant). Improved answer quality (+4 pts overall; +0.15 F1 on meta). **MLOps:** Docker + CI/CD; reproducible experiments/registries; basic K8s. **LLM inference:** batching, prompt caching, low-precision weights (4/8-bit) for latency/cost.

Core Skills

LLMs/RAG: retrieval, prompt engineering, context templates, **Graph RAG**, offline eval.
NLP: entity extraction/classification, normalization/post-processing, RU/EN corpora.
ML: supervised/unsupervised, metrics (F1, precision/recall), error analysis.
MLOps: Docker; *CI/CD (GitHub Actions)*; experiment tracking/model registries (e.g. MLflow/DVC); basic Kubernetes; monitoring/dashboards.
Stack: Python (PyTorch, scikit-learn, numpy, pandas, matplotlib), SQL, C++, Kafka, Hadoop; Vector DBs (FAISS, Qdrant).

Experience

Sberbank — IDP (Intelligent Document Processing) *Data Scientist / ML Engineer (NLP)* 10.2024–07.2025

- Built internal **RAG** & **Data Enrichment** services: retrieval → templates → extraction/validation.
- Designed and shipped **Graph RAG** (meta-chunks; cross-doc edges via section proximity/co-occurrence); vector search via **FAISS/Qdrant**.
- Achieved **+4 pts** overall quality and **+0.15 F1** on meta-questions (BERTScore/F1 harness; regression tests).
- **MLOps:** packaged inference as Dockerized API; *CI/CD (GitHub Actions)* for tests/build/deploy; latency/error dashboards; experiment tracking.
- **LLM inference optimizations:** batch generation, prompt caching, low-precision weights (4/8-bit) to reduce p95 latency and cost.

Applied Research Laboratory, VNIIA (Rosatom) *Python/C++ Dev & Algorithm Engineer* 09.2022–10.2024

- Algorithms in Python/C++; practical ML/NLP within research workflows (NDA).
- Data pipelines and evaluation utilities; emphasis on reproducibility and code quality.

Tinkoff — Project Internship *Analyst* 06.2022–07.2022

- Data analysis/reporting; SQL optimization; delivered team case solution.

Selected Projects

Marketplace Intelligence — AI RAG database for marketplace analytics: ingestion (scrapers/CSV/API) → cleaning & chunking → embeddings → vector DB (**Qdrant/FAISS**) → RAG (price/stock changes, seller QC, summarized insights); Dockerized API; eval: recall@k, latency.
DraftEdge — Dota 2 match outcome predictor (01.2024–02.2024): features on drafts/player stats; baseline ML; CV with F1.

Education

MIPT — M.Sc. (in progress), Applied Math & CS 2024–2025
HSE University — B.Sc., Physics (Quantum Computing) 2020–2024
Additional: Deep Learning School (2024); Yandex Algorithm Trainings (2024); Samsung CV (2024); Vega Institute (2025).

Languages

English — B2 | **Russian** — Native | **Japanese** — N3 | **Chinese** — A1