Aleksandr Shirokov

Remote | shirokov.aleksandr19@gmail.com | +7 (950) 015-06-40 | Eng_{B2} || $\text{Rus}_{\text{NATIVE}}$

aleksandr-shirokov | aptmess | aptmess.io

Summary

T3 Fullstack AI Software Engineer with more than **5** years experience with **Python** as the main programming language, **Data / MLOps** skills and **experience to launch AI Products from zero to production**. **Graduate** of the *master's degree* in **ITMO University** (Saint-Petersburg) on *Big Data and Machine Learning* specialization, **Moscow State University (MSU)** *AI Masters* **graduate** on *Data Science and Data Engineering* specialization, **Yandex and Tinkoff** *Backend academy* **alumni**, machine learning competitions participant.

Education

ITMO University — *Sept 2021 – June 2023*
MS in Big Data and Machine Learning

- GPA: 5.0/5.0
- Diploma Publication: Cycle Generation Networks for Sign Language Translation

Moscow State University - AI Masters — *Sept 2020 – June 2023*
SPE in Data Engineering and Data Science

- GPA: 4.65/5.0

Yandex & Tinkoff Backend Academy — *Jan 2022 – Sept 2022*
Additional Courses: Python Backend, Databases, SRE, DevOps

- Certificate

Saint Petersburg State University of Economics — *Sept 2017 – June 2021*
BS in Applied Math and Computer Science

- GPA: 4.71/5.0
- Diploma Publication: Correction of spelling errors and typos in the text using BERT

Physical and Mathematical Lyceum 239 — *Sept 2010 - June 2017*
High School Graduate

- GPA: 4.5/5.0

Experience

Team Lead MLOps Engineer — *Remote*
Wildberries — *May 2024 – ...*

- **Role and responsibilities -** Lead MLOps Engineer in a team of 6 **developers** for 5 **MLOps Streams**: RecSys Products, Pipeline Orchestration, Online Inference DL/LLM, ML Tracking and ML Tools
- **My Team and I launched** and participated in **ML System Design** for many **RecSys Business Products with ML**, which **increased revenue of RecSys Team to third of total revenue of WB**:
  - **Developed and Launched** product Search by Photo V2 using deployed by our team embedding database *Qdrant*, **daily new embeddings updates** using deployed by our team *Airflow* and *Triton* for inference, that **increased the revenue** of product 4 *times*;
  - **Developed and Launched** product Autogenerated description of product WB card using Mixtral7bx8 in vLLM on 50.000 **sellers** - this feature will be **monetized** in future;
  - **Created Triton Instance** with **HPA** and **daily model update with zero downtime** for Nearline calculation of User Embeddings in **8000 RPS**;
  - **Developed and created** Dags in *Airflow* with difficult *business logic* and large amount of steps, integration connections (more than 15) and MIG for **Item2Item Recs**

- **My Team and I launched** lots of releases for **MLOps Infrastructure**:
  - **Made an major release** for **Python** library MLTool 1.0.0 with 97% **coverage**, **automatic and versioned documentation** for library by code and lots of useful features - wrappers for Triton, Airflow, DB Connectors, S3Client[s5cmd], Docker Container Fixtures for integration tests and Quick Insert to Postgres. More than **250 developers** are using library and love it for **clean code** and **candy features**. We also **created a survey for developers** to **popularize MLTool** in quiz-mode - *How well do you know MLTOOL?*
  - **Created library** Bertolt for DL utilies, such as *Model weights conversion: PyTorch, Onnx, OpenVino, TensorRT*. This feature **significantly decreased** *Time-To-Market* for **launching DL** models in *Triton* and **systematised** the *process of deploying in production*;
  - **Developed and launched** Airflow in *K8s* as Pipeline Orchestrator with main killer feature - **launching DAGS in different K8s clusters**. **Started** process of DAGs migration to *Airflow* from *Prefect*;
  - **Patched** open-source code for fileserver in *ClearML* to make proxy to **S3** for artifacts instead of local storage, also **created automatic user creation**. **Prepared** *ClearML* to launch in production.
- **Made what Team Lead shoud do**: **made comfortable planning process**, **generated new ideas for tasks**, **participated** in tones of **interviews for new** MLOps developers, **launched Tech Demo** inside MLOps Team, **participated** in code review, demo's, retro's and grooming, **wrote digest's** of sprint results

T3 MLOps/Inference Engineer *Remote*
Wildberries *May 2022 – May 2024*

- **Developed, optimized and deployed** more than **50 ML Pipelines/Web Services** for 7 product teams with *full oriented software development lifecycle* (testing, monitoring, alerting, tracing), that **highly increased the revenue** of Recsys Department(*Personal Recs, Visual Similar Products, Item2Item Recs, Matching*), using **self-created pipeline of orchestration** using *K8s Cronjob + Prefect*;
- **First in company integrated** Triton Nvidia Inference Server as the standart-de-facto technology for *DL inference*. **Launched 3 business products** with Product Teams (*WB Lens*, *Search by Photo*, *Automatic filling of product card attributes*), that uses this technology with *OnnxRuntime* & *TensorRT* backends, that **decreased the latency** of services and **optimized replication** of servises and **GPU utilization**;
- **First in company deployed** the largest open source **LLM Mixtral7bx8** using *TensorRT LLM Backend* and Distributed Queue Service with *Celery + Redis + RabbitMQ* for first version of Product: *Autogenerated description of product WB card*;
- **Created** Python library MLTool for common DS tasks implementations with **96% test coverage** and documentation - **integrated** in more than **60 projects** and **decreased** *Time-To-Market* for MLE developers;
- **Created** *Machine Learning Repository Template* that allowed DS to deploy their ML Pipelines **in less than 10 minutes (!)**, **significantly decreased** *Time-To-Market* and **set** *Unified Code Style* for each production pipeline and research;
- **Deployed** MLOps open-source technologies on *K8s Cluster's* using *Helm charts*: *Milvus-On-Cluster*, *Prefect*, *Label Studio ML Backend* and others, **developed handy dashboards** with *monitoring and alerting* for each of deployed technology in *Grafana*;
- **Launched technical documentation** using MKDocs Material for our *MLOps Team* - **wrote** about processes, best practices, technologies usage's examples more than **15000** lines of useful information;
- **Participated in planning**, demo's, **code reviewing**, **mentoring** for Juniors, worked in *small MLOps Team* (2 developers), **created** with my Head fundamental MLOps infrastructure from **zero**!

T1 Big Data Engineer *Remote*
Grid Dynamics *Jan 2022 – April 2022*

- Studied cloud technologies for promotion to the role of T2 Big Data Engineer, took advanced classification courses and courses for obtaining certificates from Microsoft and AWS.
- **Reason to leave:** Grid Dynamics closed offices in Russia, unable to relocate.
- **Technologies:** Python 3.x, Amazon S3, Amazon DynamoDB, PySpark, Airflow, GCP

Career Start Path: Intern → T1 Data Engineer *Saint Petersburg, Remote*
Adhack.io →SkillFactory →JetBrains →4People → GreenAtom *July 2019 – Dec 2021*

Projects

Doyeshka: Web App Assistant to minimize amount of expired products in you fridge *May 2024* —

- **Lead and developed a web app** with a team allowing users to save info about *bought products* and **get alerts** before their *products inspire*
- *Tools Used*: FastAPI, PDM, Python3, SQLAlchemy, Docker, PyTest, React, TypeScript, Prefect

Stock Price Forecast based on News Context: *June 2022* — *Gitlab*

- **Lead Team and developed** an project for Big Data course work with **modern stack** of technologies
- *Tools Used*: Kafka in K8s, Distributed ClickHouse on Cluster, PySpark ML, Python3, Scrapy, Tableau

Backend service for Yandex Price on FastAPI: *July 2022* — *Github*

- **Developed an Backend** service for Yandex Price Intern on FastAPI by input swagger info - passed the exam
- *Tools Used*: FastAPI, Poetry, Python3, SQLAlchemy, Docker, PyTest

Competition Achievements

- **YaProfi 2022-2023** *Software Engineering* - Winner x2
- **AI Journey Contest 2023** - AI4Biology Silver Medal (nelepie team)
- **Digital Breakthrough 2022** *Predict popularity of news* - Bronze Medal, Money prize
- **Digital Breakthrough 2021** *Missing Planes Founder* - Bronze Medal
- **Data Fusion Contest 2021** *Goodsification* - Bronze medal (nelepie team)
- **GPN Intelligence CUP 2020** *Data Engineering* - 1st place

Technical Skills

Languages: Python3, Rust, SQL, Wolfram-Mathematica, LaTeX.

Databases & Brokers: PostgreSQL, ClickHouse, Greenplum, Redis, S3[s5cmd], MongoDB, Kafka, RabbitMQ

Devops: Docker, Docker-Compose, K8s, K9s, Helm Charts, Gitlab CI/CD, HPA, Multi Instance GPU, Vault

SRE: Prometheus[Thanos], ELK, Grafana, KSM, AlertManager

Big Data: Hadoop, HDFS, PySpark, Apache Zeppelin, AWS stack, Parquet

Python 3.x:

- **Dependency Management:** ASDF, PDM, Poetry, Rye, uv, Nexus
- **Backend:** FastAPI, SQLAlchemy, Databases, Pydantic, Httpx, alembic, Jinja2, Bootstrap, Loguru, Celery, Streamlit
- **Linters / Documentation:** Mypy, Black, Isort, Pylint, Flake8, Commitizen, Ruff, MkDocs-Material, Copier
- **Testing:** PyTest [coverage, mock, asyncio], Nox, Pytest Coverage, Deptry, Unit & Integration tests, Docker Containers for Testing

MLOps:

- **Data Orchestrators:** Airflow, Dagster, Gitlab Schedules, Prefect, K8S CronJob
- **ML Tracking / Research:** DVC, MLFlow, ClearML, Label Studio, Jupyter Hub
- **Online Inference:** Triton Nvidia Inference Server, ONNXRuntime, OpenVino, TensorRT
- **LLM Inference:** vLLM(Mixtral7bx8, Gemma), TGI, lmdeploy, Triton TensorRT LLM Backend, LLM Benchmark, Text Embedding Inference, LLaVa Deployment
- **Embedding Database:** Qdrant, Milvus-On-Cluster
- **ML Engineering:** Polars, LightGBM, CatBoost, Optuna, SHAP, PyTorch, TorchVision, Albumentation, accelerate, Transformers, Datasets, Deeppavlov

Platforms: Mac OS X, Ubuntu 22.04. **Issue Tracking**: YouTrack.