GitHub: Slavikss
Telegram: sslava$_g$
LinkedIn: vguch
Diplomas

Vyacheslav Guch

vyacheslavguch@gmail.com

Education

Bachelor in Computer Science, HSE University **2023-2027**
Applied Mathematics and Informatics, ***State-funded***

Main courses: Linear Algebra, Calculus, Python programming, Computer architecture, ML/DL, Data Structures, Data Analysis etc.

Efficient DL Systems **2024**
YSDA and HSE course

Worked with ONNX, DVC, model quantization, experiment tracking via MLFlow, wandb and dvc, made efficient deploy GPT2 with Docker, Grafana. Finetuned 3B model via Model-parallel. Optimized inference with OpenVino.

DL Courses **2022**
- DLS : Introduction to DL course. Successfully completed the tasks based on CNN's: Image classification, Image segmentation, Image classification
- **Tinkoff DL/DL Advanced** (link): Succesfully completed labs and mini-projects with Neural Networks: Transformer models, Metric Learning, Language Modelling, RNN, AE/VAE, GAN, Segmentation, Translation, Effective Learning and Deploy
- All my 10+ SQL, AI, Programming competitions and courses diplomas: ->link <-

Experiences

MLE/ Co-founder, Shperling AI **Apr 2024 – May 2025**
Built enterprise-grade RAG solutions with 95%+ accuracy and 25s response time

- Led and mentored a team of 5 ML engineers in building production RAG system: guided architecture decisions for Qwen-2.5-32B-Instruct-GPTQ-Int4 (local) and GPT-4o-mini (API) integration, achieving 92.4% context recall and 87.3% precision on RAGAS benchmarks. Coached team on vector store optimization with Milvus and BM25 hybrid search.
- Established best practices and trained team on secure text2SQL development: implemented code review process focusing on SQL injection prevention and query validation, resulting in 82% accuracy on complex queries (GROUP BY: 91.4%, JOIN: 82.9%) and 87.5% Russian language support.
- Designed and supervised implementation of microservices architecture, coaching junior developers on:
  - Production-grade FastAPI development and Docker containerization
  - Document processing pipelines with chunking optimization (rag-pilot)
  - Secure database integration and query validation (text2sql)
  - External API integration achieving 92% relevance (funccal)
  - Distributed storage systems with Milvus, MinIO, and ETCD
- Introduced MLOps practices through hands-on workshops: trained team on MLflow experiment tracking, W&B monitoring, and CI/CD pipeline development. Reduced deployment issues by 40% through systematic code review and documentation.

BigData Analyst, Beeline Kazakhstan **2023**
Beeline is the largest Fintech company in Kazakhstan

- Interacted with production DB's based on HDFS, PySpark, Kafka, PostgreSQL to upload and send data.
- Developed and brought to production SQL triggers for MyBeeline App.
- Refactored Multisim Model, increased performance for 30 percents(F1 score)
- Communicated with customers, promptly and succesfully solved existing and new issues.

DANO Olympiad Winner, HSE and Tinkoff **2022**
Data Analysis National Olympiad

- Solved different analytical, data analysis tasks
- Led team of 5 people, which made research on dataset of movie ticket sales in 2020
- Organized team and tracked a progress in Miro, Notion
- Made EDA, Feature engineering, Data vizualizaon using pandas/seaborn etc.

Totally - **23rd/7k+ place**(individual ranking), **9th/56 place**(team tour)

Projects

GPT2 Pretraining from Scratch **2024**
- Implemented vanilla GPT2 on pure PyTorch.
- Optimized with flash attention, compiling, autocasting, right model initialization, mixed precision
- Used gradient accumulation for 0.5M batchsize effect. Trained on FineWebEDU 10B dataset with a DDP training on 2x RTX4090
- Achieved 32 percent HellaSwag score on my model vs 28 percent on original model

XmasHack Winner **2024**
- Took 1st/10 place in a hackathon aimed at optimizing marketing expenses of the largest CIS TV channel via TVR Index prediction on historical tabular data
- Used Optuna, Catbooost, Complicated Feature Engineering with seasonality detection
- Achieved 23 percent MAPE at private test set. Closest to us solution had 25 percent MAPE