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

ML Tech Lead / Co-founder, Shperling AI **Apr 2024 – May 2025**
Startup: On-premise RAG solutions for enterprise clients

- Architected and deployed three on-premise Retrieval-Augmented Generation (RAG) systems for enterprise clients (internal support, Swiss consulting, fintech with 1M+ users):
  - For internal support, implemented RAG pipeline with Qwen-32B as the core LLM and RAGAS for automated evaluation; achieved 95% accuracy on internal metrics and robust performance on golden set benchmarks.
  - All RAG deployments included strict SQL injection detection, safe database connection management, and query rate limiting for production safety.
- Developed and productionized a text2SQL analytics pipeline for B2B retail:
  - Used Qwen-32B for SQL generation, validated outputs on a golden set, and reduced analytics latency from 8h to 1h.
  - Integrated SQL injection detection and secure DB connection logic to ensure safe execution in enterprise environments.
- Built and managed an RD team (5 junior ML engineers): organized Scrum sprints, technical reviews, and delivered 5+ custom RAG prototypes, one of which became the commercial product core.

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