Stanislav Ilyushin
https://github.com/StanislavII
Email : StanislavIlyushin@mail.ru
Mobile : +7 (965) 102-13-02

EDUCATION

Higher School of Economics Moscow, Russia
Master's Degree. Faculty of computer science. 2021-2024
Data Science. Institute of Information Transmission Problems RAS; GPA: 8.2

PROFFESSIONAL EXPERIENCE

Huawei Moscow, Russia
Senior Engineer 08/2024-

- Delivered lossless inference acceleration for Llama3 and Qwen2.5 using an EAGLE-based draft–model hybrid system. Published in the READER paper. Achieved up to **5× latency reduction** over autoregressive baselines, surpassing EAGLE3 state-of-the-art performance.
- Built and deployed **GLDS**, an image token pruning algorithm for Qwen2.5-VL, enabling scalable large-batch inference. Results published in GLDS. Delivered up to **1.5× throughput improvement** with under **1%** quality loss.
- Implemented a speculative Jacobi decoding pipeline with VQ-VAE–based verification relaxation, achieving approximately **2× inference speed-up** for production image-to-text autoregressive models.
- Accelerated Qwen3-VL and Qwen3-VL MoE inference by integrating an optimized **GPrune** algorithm into `vLLM` and `vLLM-Ascend`. Improved end-to-end throughput by up to **1.3**×, including CUDA graph capture–based inference.
- Designed and implemented a speculative decoding system using a diffusion-based draft model for the `sglang` framework, outperforming autoregressive baselines by up to **1.25**× in high-concurrency serving environments.

Sberbank Moscow, Russia
NLP engineer 08/2023- 08/2024

- Developed a multi-agent system using graph transitions based on llama for Gigachain. Developed Gigachat summarization abilities.
- Gigachat study based on Prompt Optimization via Adversarial In-Context Learning approach.

Rostelecome Moscow, Russia
Middle Data Scientist. Machine Learning Department of the Credit Risk Center 03/2020- 08/2023

- Developed a full stack system for classifying income-generating contracts by risk level based on their description.
- Developed a model for clustering conversations between a robot operator and a debtor based on graph models such as watset and Growing Neural Gas.

Lectures Moscow, Russia
Ranepa 01/2024 – 06/2025

- Applied NLP Optimization

OTUS 06/2025 – 01/2026

- NLP. Advanced

ACHIEVEMENTS

Prize-winning place at HSE Higher League Student Olympiad. 02/2021
Discipline: Mathematical methods of economic analysis.

Finalist x3 of the Olympiad "I am a Professional" in mathematics. 02/2020-02/2023

ADDITIONAL ACTIVITIES

Public performance: Discussion and demonstration of practical examples of the use of artificial intelligence to solve business problems.

- Speech at a hackathon on artificial intelligence as a speaker on the topic "Retrieval Rerank LLM". November 23, 2023.
- Speech at the conference "One step ahead!"Global CIO in the artificial intelligence section with the topic of graph autoencoders. October 5, 2023.
- Speech at the 17th annual conference "Moscow Evenings"in the section "Neural networks: practical application"from Rostelecom. September 24-27, 2023.

Skills

IT: Python (pytorch, transformers), Docker, Airflow, Linux, SQL, LaTeX

Languages: English (Upper - Intermediate), Russian (Native)

Interests

Science interests: NLP, Numerical Linear Algebra

Hobby: Hockey, Music