Vladislav Kruglikov

Senior LLM Inference Research Engineer at T-Bank; formerly Yandex

vladislavkruglikov.com *vladislavkruglikov@icloud.com* *linkedin.com/in/vladislavkruglikov*

Summary

LLM inference and infrastructure research engineer with 4+ years of experience in speculative decoding, KV cache efficiency, serving, and training-aware optimization. Led inference, model training, and production infrastructure projects improving throughput, latency, and model delivery speed. Author of a technical blog at vladislavkruglikov.com on efficient inference and training, speculative decoding, serving tradeoffs, and practical systems work. Served as a technical interviewer for algorithms, data structures, and NLP roles.

Experience

T-Bank *Russia's leading digital bank, 51M+ customers* — *Moscow, Russia*
Senior Research Engineer, Core LLM Team — *March 2026 – Present*

- Driving training-aware LLM inference optimizations to improve the cost efficiency of open-source models.

Yandex *Russia's largest IT company, 100M+ MAU* — *Moscow, Russia*
Senior Research Engineer, YandexGPT — *July 2025 – February 2026*

- Researched and implemented long-context KV cache pruning algorithms, achieving superior efficiency over open source alternatives.

T-Bank *Russia's leading digital bank, 51M+ customers* — *Moscow, Russia*
Senior Research Engineer, Core LLM Team — *February 2022 – July 2025*

- Led the inference team to build and open-source **T-Pro 2.0 EAGLE**, a speculative decoding draft model for T-Pro 2.0. Built the training pipeline from scratch and delivered substantial throughput and latency improvements. Released publicly on Hugging Face: t-tech/T-pro-it-2.0-eagle.
- Served on the **technical interview committee**, evaluating candidates in algorithms, data structures, and natural language processing.
- Led a team to build a model fine-tuning and inference platform that dramatically improved trained-model time-to-market by leveraging PEFT and an efficient inference runtime.

Publications

T-pro 2.0: An Efficient Russian Hybrid-Reasoning Model and Playground — *Moscow, Russia*
Proceedings of EACL 2026 — System Demonstrations Track — *March 2026*

Selected Writing

Speculative Decoding — *Personal Blog*
Technical article on draft systems, verification, and serving tradeoffs — *May 2026*

Education

Higher School of Economics — *Moscow, Russia*
B.Sc. in Applied Mathematics and CS, Major in Distributed Systems — *August 2022 – June 2026*

Accomplishments

Winner, NLP Hackathon by VTB Bank — *Moscow, Russia*
*Led the team to **1st place** by building a personalized news feed recommender system* — *October 2022*