

Sergei Ermolaev

👉 @J1nsei 📩 saermolaev.w@gmail.com ☎ +7 (930) 802-31-25 ⬇ Nizhny Novgorod (UTC+3) RU CV

Summary

Machine Learning Engineer (GenAI/LLM), **2.5 years** of end-to-end production ML experience: RAG, LLM inference, multi-label classification, hybrid search, CV tracking; Education — HSE University (AMI + Data Mining).

Experience

Proton (IT company of "AGAT" Group, top-5 Russian car dealers ↗, 120+ showrooms)

Nizhny Novgorod
Nov 2023 – Present

Machine Learning Engineer:

- Multilabel call classification: **F1-samples 79% → 94%, EMR 3% → 61.6%**; sub-second CPU inference, >1000 daily calls; business impact ~\$1.7M. Data collection and processing: transcription, cleaning, analysis, stratification in a multilabel setting; conducting experiments, production integration.
- RAG Assistant** (RAGFlow + vLLM): **user satisfaction 4.7/5, TTFT p99 4.6s** under load with long context, **faithfulness 95%**. Customized framework functionality for corporate Knowledge Base specifics, stack/model selection, custom document correction and parsing pipeline, Docker, production deployment with monitoring and tracing, performance and quality benchmarking, user and support documentation.
- Car parts search system: reduced manual requests **300→20/day** (-93%). Database selection based on built-in capabilities and speed. Implemented **hybrid search**, custom ranking function and thresholds, wrapped in FastAPI + Docker; created abbreviation dictionary (regex + LLM + manual editing) and documentation, integration with production server.
- LLM call labeling** (Mistral/Qwen, zero-/few-shot, **LoRA Qwen2.5-3B**): Accuracy up to 90% (LoRA), production — XLinear 88% (CPU online/offline).
- Test drive **CV tracking**: YOLOv5, multithreading, crops; accuracy 65%→87%, up to 15 fps on Consumer GPU.
- LLM API for **summarization**, car description generation, and arbitrary prompts.
- Ticket routing **Agent** (GANDIVA). Agent prototype on RAGFlow with structured output checks (JSON recovery, LLM verification).

Education

HSE University

Master's degree in Data Mining

Nizhny Novgorod, Russia

Sept 2022 – June 2024

HSE University

Bachelor's degree in Applied Mathematics and Informatics

Nizhny Novgorod, Russia

Sept 2018 – June 2022

Professional Development

Yandex School of Data Analysis (SHAD) Intensive: **LLM Scaling Week ↗**

Nov 10, 2025 – Nov 14, 2025

Skills

DL/ML: Python, NumPy, RAGFlow, PyTorch, transformers, vLLM, llama.cpp, deepeval, mineru, prompt engineering, infinity, unsloth, PECOS, scikit-learn, WhisperX, spaCy, ultralytics

Infra/Tools: Linux, Docker, FastAPI, Git, DVC, streamlit, LangFuse, W&B, matplotlib, Grafana, Kubeflow

Data: pandas, polars, SQL

Languages: Russian (Native), English (Intermediate)