Сергей Ермолаев

@J1nsei saermolaev.w@gmail.com +7 (930) 802-31-25 Нижний Новгород ENG CV.

Обо мне

Machine Learning Engineer (GenAI/LLM), **2.5 года** продакшн-ML end-to-end: RAG, LLM-инференс, мультилейбл-классификация, гибридный поиск, CV-трекинг; образование - ВШЭ (ПМИ + ИАД).

Опыт

Протон *(IT-компания ГК* ***"АГАТ"****, топ-5 автодилеров РФ, 120+ автосалонов)* Нижний Новгород
Machine Learning Engineer: Ноябрь 2023 – Н. В.

- Multilabel-классификация звонков: **F1-samples 79% → 94%**, **EMR 3% → 61.6%**; CPU инференс за доли секунды, >1000 звонков ежедневно; эффект **~170 млн ₽**. Сбор и обработка данных: транскрибация, очистка, анализ, стратификация в условиях многометочности; проведение экспериментов, интеграция в прод.
- **RAG**-ассистент (RAGFlow + vLLM): **user satisfaction 4.7/5**, **TTFT p99 4.6s** под нагрузкой с большим контекстом, **faithfulness 95%**. Доработка функционала библиотеки под особенности корпоративной БЗ, выбор стеков/моделей, кастомный пайплайн исправления и парсинга документов, Docker, разворачивание проекта на прод-сервер с мониторингом и трейсингом, измерение производительности и качества, подготовка документации по использованию и поддержке.
- Система подбора автозапчастей: снижение обращений **300→20/день** (-93%). Подбор БД по встроенным возможностям и скорости работы. Реализация **гибридного поиска**, кастомная функция ранжирования и пороги, оборачивание в FastAPI + Docker; создание словаря сокращений (regex + LLM + ручная вёрстка) и документации, интеграция с прод-сервером.
- **LLM-разметка** звонков (Mistral/Qwen, zero-/few-shot, **LoRA Qwen2.5-3B**): Accuracy до 90% (LoRA), прод — XLinear 88% (CPU онлайн/офлайн).
- **CV-трекинг** тест-драйвов: YOLOv5, мультипоточность, кропы; точность 65%→87%, до 15 fps на Consumer GPU.
- LLM API для **суммаризации**, генерации описаний авто и произвольных промптов.
- **Агент** маршрутизации заявок (GANDIVA). Прототип агента на RAGFlow с проверками структурированного вывода (восстановление JSON, верификация LLM).

Образование

Национальный исследовательский университет «Высшая школа экономики» Нижний Новгород, Россия
Магистратура, Интеллектуальный анализ данных *Сент. 2022 – Июнь 2024*

Национальный исследовательский университет «Высшая школа экономики» Нижний Новгород, Россия
Бакалавриат, Прикладная математика и информатика *Сент. 2018 – Июнь 2022*

Дополнительно

Интенсив ШАД Яндекс: LLM Scaling Week *10.11.2025—14.11.2025*

Навыки

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)