

АЛЕКСЕЙ ПАВЛОВ

+7(921)4396406 | Санкт-Петербург | tg: @glaucomaa | github: glaucomaa | aamocualg@gmail.com

ОПЫТ РАБОТЫ

Data Scientist/Engineer

2024 - н.в.

Best Place

- Внедрил ML-сервис категоризации товаров (*Sentence-BERT*): F1 ↑ с 74% до 88%, время обработки ↓ в 5 раз.
- Автоматизировал пайплайн категоризаций (*RuBERT*), сократив объём ручной разметки на 80 %.
- Обучил модель прогнозирования спроса (*Prophet* + *CatBoost*): MAPE ↓ с 18 % до 9 %.

ML / NLP Research Engineer

2025 – н.в.

EMS VK

- Развил открытый фреймворк для обучения и экспериментов: модульный рефакторинг, поддержка логирования, сократил используемое дисковое пространство на 60 %.
- Формулировал и приоритизировал гипотезы (одно/много-головое внимание, схемы Local-Global, варианты RoPE); готовил план тестов совместно с PI.
- Создавал и поддерживал конфигурации экспериментов (Hydra / YAML / JSON / Bash); автоматизировал массовые абляционные испытания на GLUE и LRA - accuracy LRA ↑ 1.5 п.п.

ПРОЕКТЫ

Потоковая ETL-платформа для управления сущностями

- Спроектировал контейнеризованную архитектуру (PostgreSQL + Kafka-кластер из 3 брокеров + Cassandra + Spark) с Docker Compose и персистентными томами.
- Реализовал REST API на FastAPI с асинхронным пулом соединений; повысил пропускную способность до 3 000 запросов в секунду.
- Собрал надёжный ETL-конвейер: атомарно сохраняю в формате JSON, стримлю данные через Kafka и записываю их в Cassandra.

RAG vs LoRA

- Разработал RAG-систему на базе *Falcon-1B* с динамическим поиском контекста, повысив среднее качество генерации текста на +27 % по 8 метрикам (BLEU, ROUGE-1/2/L, METEOR, Precision/Recall/F1).
- Снизил требования к видеопамети в **3 раза** через QLoRA при сохранении >90% качества.
- Создал бенчмарк на **12 000 QA-примерах** из 3 датасетов, подтвердив эффективность RAG-подхода для задач вопросно-ответных систем.
- Предложил и сравнил **два алгоритма сегментации**, показав превосходство первого над статическими подходами на 2 п.п. и отставание от идеального подхода всего на 7 п.п.

ОБРАЗОВАНИЕ

Высшая школа экономики, Санкт-Петербург (магистратура),

Школа физико-математических и компьютерных наук,

Машинное обучение и анализ данных

2024 - 2026

Средний балл: 8.5/10.0

Санкт-Петербургский государственный университет, Санкт-Петербург (бакалавриат),

Факультет прикладной математики - процессов управления, ФИИТ,

Программирование и информационные технологии

2020 - 2024

Средний балл: 4.76/5.00

ПУБЛИКАЦИИ

- Methods of Fine-Tuning: RAG vs LoRA - конференция CPS'24, HIT, Harbin

НАВЫКИ

Python, C/C++, Rust, PyTorch, Transformers, scikit-learn, CatBoost, ONNX Runtime,
Docker, FastAPI, Kafka, ClearML, Hydra, Linux, Git, SQL, LoRA, RAG, Quantization, Pruning