



Третьяков Андрей Андреевич

Мужчина, 23 года, родился 21 сентября 2002

+7 (917) 5482432

tret'yakov.xrf@gmail.com — предпочитаемый способ связи

telegram: @xrayfun

xrayfun.ru: <https://xrayfun.ru>

Проживает: Москва, м. Новогиреево

Гражданство: Россия, есть разрешение на работу: Россия

Не готов к переезду, готов к командировкам

Желаемая должность и зарплата

Lead LLM Engineer / Руководитель направления ML (RAG/Agents/NLP)

420 000 ₽ на руки

Специализации:

- Дата-сайентист
- Другое
- Программист, разработчик
- Руководитель группы разработки

Тип занятости: полная занятость

Формат работы: на месте работодателя, удалённо, гибрид

Желательное время в пути до работы: не имеет значения

Опыт работы — 6 лет 11 месяцев

Апрель 2026 —
настоящее время
5 месяцев

Альфа-Банк (Россия)

www.alfabank.ru

Финансовый сектор

- Банк

Ведущий разработчик

Специалист по операциям с языковыми моделями - LLMOps

Август 2024 —
Апрель 2026
1 год 9 месяцев

Код безопасности

Москва, www.securitycode.ru

Информационные технологии, системная интеграция, интернет

- Разработка программного обеспечения

Специалист

Разрабатываю RAG-агента для сотрудников техподдержки. Поиск информации по корпоративной документации, формирование ответов в формате "ссылка + выдержка + итог".

Построил pipeline: обработка/парсинг документов (Docling), векторизация, хранение эмбедингов и поиск в Qdrant, метаданные и связи - PostgreSQL.

Использую reranker и дообученную (fine-tuned) модель embeddings на корпусе корпоративной документации.

Инференс LLM - облачные (OpenAI, GigaChat) + локальные модели в контуре (Ollama, vLLM) - подбор режима под ограничения on-prem/стоимости/скорости.

Docker + GitLab CI/CD, on-prem деплой; MLflow для трекинга; самописные тесты корректности

RAG и стресс-нагрузки инференса.

Open-WebUI, n8n (пользовательские цепочки, запуск сценариев, сбор фидбэка).

Параллельно: пентесты внутренних и внешних продуктов, разработка корпоративных web-приложений на Django, автоматизация внутренних процессов.

Управление в рамках отдельных проектов - декомпозиция задач, синхронизации со стейкхолдерами, контроль выполнения, отчетность.

Стек: Python, Django, Qdrant, PostgreSQL, Docling, OpenAI, GigaChat, Ollama, vLLM, MLflow, Docker, GitLab, Open-WebUI, n8n.

Январь 2023 —
Декабрь 2023
1 год

Cristalix

Java-разработчик

Разработка серверной логики/механик (Java, Gradle).

Интеграции, работа с данными, оптимизация и сопровождение функционала.

Проектирование/поддержка БД (MySQL, SQLite).

Июнь 2023 —
Июль 2023
2 месяца

Ростелеком

Россия, www.rostelecom.ru

Информационные технологии, системная интеграция, интернет

- Интернет-провайдер

Телекоммуникации, связь

- Оптоволоконная связь
- Фиксированная связь
- Мобильная связь

Специалист техподдержки (практикант)

Диагностика и поддержка рабочих мест/сетевой инфраструктуры, настройка оборудования, поддержка удалённых сотрудников.

Август 2021 —
Декабрь 2021
5 месяцев

МАГНИТ, Розничная сеть

Россия, www.magnit-info.ru

Розничная торговля

- Розничная сеть (продуктовая)
- Розничная сеть (одежда, обувь, аксессуары)
- Розничная сеть (drogerie, товары повседневного спроса)

Web-разработчик (практикант)

Разработка учебного WEB-сервиса (задачи от куратора).

Декабрь 2018 —
Август 2019
9 месяцев

Фриланс

Руководитель группы разработчиков

Ведение проектов web-разработки: коммуникация с заказчиком, декомпозиция и распределение задач, контроль выполнения, работа по Agile-подходу.

Апрель 2018 —
Декабрь 2018
9 месяцев

Фриланс

Web-разработчик

Python backend (Django), базы данных, web-разработка.

Апрель 2016 —
Ноябрь 2018

GameCloud

2 года 8 месяцев

Java-разработчик

Разработка решений на Java (в т.ч. под игровые серверы), участие в командной разработке.

Образование

Бакалавр

2024
Бакалавр

Национальный исследовательский университет «МЭИ», Москва
Инженерно-экономический институт, Информационная безопасность

Повышение квалификации, курсы

2020

Курс полной подготовки и профессиональной ориентации
Московская Школа Программистов, web-разработчик

Навыки

Знание языков

Русский — Родной
Английский — B1 — Средний

Навыки

LLM RAG embeddings reranker agentic workflows prompt engineering
fine-tuning OpenAI API OpenAI GigaChat Ollama vLLM Qdrant
PostgreSQL JSON SQL Docling Whisper MLflow GitLab CI/CD
Docker on-prem deployment load testing Python Django REST API
HTTP Linux Git Java

Опыт вождения

Имеется собственный автомобиль
Права категории B, BE

Дополнительная информация

Обо мне

LLM/ML-инженер с backend-базой (Python/Django) и практикой в информационной безопасности (пентесты).

Собрал и внедрию RAG-агента для техподдержки - поиск ответов по корпоративной документации и тикетам, reranking, адаптация embeddings на внутреннем корпусе.

Работаю с облачными LLM (OpenAI, GigaChat) и локальными (через Ollama/vLLM) в on-prem контуре. Работаю над несколькими pet-проектами с LLM.

Для деплоя - Docker + GitLab CI/CD; трекинг - MLflow; качество - автотесты корректности RAG + стресс/нагрузочные тесты инференса.

Интересуют роли LLM Engineer / Applied LLM (RAG/Agents), где можно доводить решения до эксплуатации и масштаба.