

Бондарев Ярослав

Telegram: @bozyk44 | bozyk44@gmail.com | linkedin.com/in/yaroslav-bondarev-84571a168 | github.com/bozyk44 | Москва, Россия | Удалённо | Русский родной, английский C1

AI Engineer (LLM / GenAI) | LangGraph · Multi-model · RAG | Python · Next.js | ASR · Eval · MLOps

Инженер с фундаментальным образованием (МГУ, факультет ВМК) и опытом 7 лет в разработке и data-инженерии, сфокусированный на построении продакшн-систем на основе LLM/GenAI. Проектирую и довожу до прода end-to-end AI-пайплайны: оркестрация нескольких LLM, RAG и гибридный поиск, ASR/голос, извлечение информации, eval-пайплайны и self-hosted inference. Сочетаю практическую AI-инженерию с системным анализом и бизнес-коммуникацией. Работал в CFRA Research (США), Яндекс.Лавке, Сбере и Fastmarkets. Английский C1.

ОПЫТ РАБОТЫ

Fastmarkets — Senior Fullstack / AI-инженер, лид разработки

Январь 2026 — наст. время

Проект: Luminarium AI — платформа оценки биржевых цен на сырьё для международного ценового агентства: транскрибация аудио -> извлечение данных -> редакторская проверка -> AI-оценка. 20 рынков, 50+ аналитиков, ~1000 звонков в месяц

Команда и процессы: 10 человек, роль — лид разработки; Scrum; взаимодействие с аналитиками, Product Owner, IT-security, Cloud Engineering / SRE и сопровождением

Обязанности

- Проектирую и развиваю end-to-end LLM-пайплайн оценки цен: от голосового ввода аналитика до готовой оценки рынка
- Разрабатываю бэкенд и фронтенд приложения целиком, оцениваю и ставлю задачи команде
- Оркестрирую мультимодельные сценарии на LangGraph.js (Gemini Flash, Claude Haiku, GPT) — маршрутизация моделей, ретраи, контроль токенов и расхода
- Интегрировал стек распознавания речи (AssemblyAI, Whisper, Azure Speech, ElevenLabs) для голосового ввода аналитиков
- Настроил observability (OpenTelemetry + LGTM), деплой в Azure, IaC на Terraform, доступ через Cloudflare Access
- Веду housekeeping команды: отслеживаю загрузку, распределяю задачи, предлагаю и внедряю оптимизации процессов
- Провожу онбординг и обучение, внедряю в команде best practices работы с AI-агентами

Достижения

- Сократил время аналитика на оценку рынка на 40%
- Разработал AI-сервис валидации наблюдений и выпустил Slack-бота (Events API) с извлечением данных по 20 рынкам
- Построил E2E eval-пайплайн и подготовил тестовые наборы для измерения качества ответов LLM по сценариям
- Пайплайн работает в прода 3 месяца, ~30 прогонов в день (прогон = звонок аналитика)

Стек: TypeScript (strict), Next.js 15, LangChain / LangGraph.js, Anthropic / Google GenAI SDK, Prisma + SQL Server, Langfuse, OpenTelemetry / LGTM, Azure, Kubernetes, ArgoCD, Terraform, GitHub Actions, Playwright, Claude Code, CodeRabbit AI

Сбер — Data Engineer

Май 2024 — Декабрь 2025

Проект: SberData — загрузка первичных данных в корпоративную аналитическую платформу банка: репликация и архивирование данных автоматизированных систем; лично обработано 30 систем-источников

Команда и процессы: 6 человек, роль — Middle Data Engineer; Kanban, Jira; координация с 10+ смежными командами

Обязанности

- Собирал и анализировал требования смежных команд, настраивал фреймворк репликации и писал скрипты загрузки
- Тестировал качество данных, вёл документацию и поддерживал self-service загрузку данных
- Настраивал дашборды и сбор данных для отчётности, писал инструменты автоматизации для self-service и сопровождения

Достижения

- Реализовал 20 реплик и архивов БД для автоматизированных систем банка суммарным объёмом 100+ ТБ
- Сократил срок подключения источника с 1 месяца до 2 недель
- Обработал 30 систем-источников, обслужил 20 разовых заказчиков загрузки в формате self-service

Стек: Hadoop (HDFS, YARN), Hive, HBase, Spark / PySpark / Spark SQL, Python (pandas, numpy), SQL, Oozie, Kafka, Parquet / ORC / Avro, Kerberos, Zookeeper, Hue, Bash

Проект: трансформация сырой выдачи модели расчёта финансовых показателей в тиражные данные и их распространение — API, фиды, PDF-отчёты, сайт. Universe 35 000 компаний: еженедельный отчёт по каждой, ежедневный — при ненулевом diff

Команда и процессы: 10 человек, Scrum, Jira; частично функции лида, самостоятельные архитектурные решения

Обязанности

- Дорабатывал пайплайн трансформации сырых расчётов модели в тиражные данные
- Развивал каналы доставки данных клиентам: API, фиды, ежедневные PDF-отчёты, публикация на сайте
- Вёл разработку и код-ревью, контролировал качество данных и доступность ресурсов
- Принимал технические и архитектурные решения по проекту

Достижения

- Ускорил обработку в 3 раза относительно legacy-решения
- Добавил 5 видов отчётов и кастомные фиды для 10 заказчиков

Стек: Python (Flask, pandas, numpy, SQLAlchemy, Alembic, pytest, boto3), PySpark, SQL, WeasyPrint, AWS (S3, Lambda, ECS, Batch, Glue, RDS, CloudWatch), Terraform, Docker, Jenkins, ETL

Яндекс.Лавка — Middle Backend Developer

Август 2022 — Август 2023

Проект: внутренняя ERP-система (Odoo) для операционных процессов сервиса в 3 странах

Команда и процессы: 8 человек, Scrum, Яндекс.Трекер и внутренняя вики

Обязанности

- Разрабатывал и поддерживал модули ERP: новые модули, кастомизация, исправление legacy-кода
- Разрабатывал интеграции с внутренними сервисами Яндекса

Достижения

- Сделал 2 интеграции с внутренними сервисами (авторизация, API)
- Участвовал во внедрении ERP в Израиле и Турции

Стек: Python, Odoo (ORM, QWeb, XML-views, queue_job), Flask, PostgreSQL, SQL, Docker, Nginx, pytest, внутренний CI Яндекса

CFRA Research, США — Junior -> Middle Backend Developer

Июнь 2019 — Июль 2022

Проект: аналитическая платформа исследовательского агентства — расчёт финансовых показателей, отчёты и дата-фиды для институциональных клиентов. Universe ~5 000 компаний

Команда и процессы: команда в США ~6 человек, рабочий язык английский; Scrum, Jira

Обязанности

- Начинал с поддержки legacy; после повышения до Middle — разработка нового функционала и переработка архитектуры legacy
- Настроил ежедневный ETL-пайплайн с контролем качества, мониторингом, алертингом и хранением промежуточных данных для observability
- Настроил расчёт показателей, ежедневную генерацию PDF-отчётов, доставку фидов и доступ к данным через API
- Переводил бизнес-требования в задачи разработки

Достижения

- Снизил число жалоб клиентов на данные на 60% после внедрения контроля качества и алертинга
- Сократил ночные вызовы on-call инженера на 40%

Стек: Python (Flask, pandas, numpy, matplotlib, SQLAlchemy, Alembic, Jinja, pytest, boto3), PySpark, SQL (Oracle, MySQL, PostgreSQL, SQL Server), XML, WeasyPrint, AWS (S3, Lambda, ECS, Batch, Glue, RDS, CloudWatch), Terraform, Docker, Jenkins, ETL

ОБРАЗОВАНИЕ

МГУ им. М. В. Ломоносова

2019

Бакалавр • Факультет вычислительной математики и кибернетики (ВМК) • Прикладная математика и информатика

НАВЫКИ

GenAI / LLM: prompt-engineering, оркестрация LangGraph / LangChain, мультимодельность (Claude, GPT, Gemini), ансамбли моделей, агентные сценарии, eval и A/B-тестирование LLM, LLM-observability (Langfuse)

RAG и поиск: гибридный retrieval (BM25 / FTS5 + плотные векторы), bge-m3, квантизация векторов int8, чанкинг с сохранением таблиц и единиц, оценка retrieval (recall@k), проверка ответа на соответствие источнику

Inference и MLOps: vLLM, AWQ-квантизация, оркестрация GPU в облаке, оптимизация KV-cache / VRAM, инженерия стоимости

ASR / голос: Whisper / WhisperX, faster-whisper, диаризация NeMo и pyannote, AssemblyAI, ElevenLabs

Языки и backend: Python, TypeScript, SQL, Bash, FastAPI, Flask, Spark / PySpark / Spark SQL, pandas, numpy, SQLAlchemy, Alembic, pytest, ETL, Prisma, Hadoop (HDFS, YARN), Hive, HBase, Kafka, Oozie, Parquet / ORC / Avro

Web и инфраструктура: Next.js, React, Azure, AWS (S3, Lambda, ECS, Batch, Glue, RDS, CloudWatch), Terraform, Docker, Kubernetes, ArgoCD, Nginx, Cloudflare, CI/CD (GitHub Actions, Jenkins), OpenTelemetry / LGTM, Playwright, Linux

Базы данных и AI-assisted разработка: PostgreSQL, SQL Server, MySQL, Oracle, SQLite; Claude Code, Codex, CodeRabbit AI

ПРОЕКТЫ (ЛИЧНЫЕ AI/ML)

rag-ru-reports — RAG по годовым отчётам российских эмитентов · демо rag-ru-reports.pages.dev · код github.com/bozyk44/rag-ru-reports

- Retrieval выбран замером на 68 вопросах и 16 конфигурациях: BM25 (FTS5) -> топ-50 -> косинус int8 даёт recall@5 80% при в 97 раз меньшем объёме вычислений; квантизация int8 сжала индекс до 4.68 МБ, медиана запроса 3 мс, грунтованность чисел 96%

ClipsForge — LLM-пайплайн: многочасовой стрим -> 8–15 готовых к публикации клипов

- Ансамбли Claude для детекции моментов, ASR на WhisperX / faster-whisper (до 16.4x real-time), self-hosted vLLM (Gemma 31B AWQ) на облачных GPU с расчётом стоимости прогона