

Коробов Евгений Владимирович

AI / LLM Engineer · агентные системы, RAG, Python

Ростов-на-Дону · удалённо, UTC+3 · в Ростове возможен гибрид или офис

+7 909 433-13-12 · ek@eko-res.ru · t.me/korobovevgen

[ai-producing.com](<https://ai-producing.com>) — полное резюме и проекты · neiroaktiv.ru

КОРОТКО

Два года строю LLM-системы и вывожу их в прод. 43 проекта, ≈400 000 строк кода, 4 Telegram-бота в постоянной эксплуатации, 6+ боевых серверов в РФ, ЕС и США, production RAG на векторной базе, мультипровайдерный роутинг с автоматическим fallback.

До этого 20 лет отвечал за процессы, которые теперь автоматизирую: контакт-центр на 350 человек в телекоме, бэк-офис на 50 сотрудников с нуля, три собственные B2B-компании. Понимаю не только как собрать агента, но какие метрики он должен двигать и что в нём сломается на втором месяце эксплуатации.

ЧТО ДЕЛАЛ В LLM-ИНЖЕНЕРИИ

RAG-контур в проде, собран с нуля. Полный конвейер: ingest → чанкинг → эмбединги → векторное хранилище → query-слой → HTTP-сервер поиска. Мультиязычные эмбединги `intfloat/multilingual-e5-base` через sentence-transformers с корректными префиксами `query:`/`passage:` и нормализацией векторов. Чанкинг 1200 символов с перекрытием 33% — размер подобран под структуру корпуса. Инкрементальная переиндексация по `source\_file`: устаревшие чанки удаляются до вставки новых, живой индекс не ломается. Метаданные и фильтрация по коллекциям. Источники: расшифровки аудио через Whisper, таблицы, документы.

Отказоустойчивость LLM-провайдеров. Роутинг между Anthropic, Google Vertex, OpenAI, Groq, Cerebras, DeepSeek, OpenRouter и Ollama с приоритетами и автоматическим fallback по таймаутам и кодам ошибок. Конфигурация моделей и приоритетов хранится в базе данных, а не в коде — провайдер меняется и переупорядочивается без деплоя. Отдельно решена практическая проблема моделей с thinking-токенами, которые съедают лимит и обрывают длинные JSON-ответы.

Проверено боем: когда основной провайдер заблокировали без предупреждения, все зависимые сервисы переведены на резервные модели за один день без простоя. Контур деградации был спроектирован до инцидента, а не после.

Агентные контуры. Долговременная память между сессиями — сквозной профиль пользователя: накопленные факты, история обращений, контекст прошлых диалогов; агент помнит клиента спустя месяцы. Решена эксплуатационная часть: рост профиля не съедает контекст и не размывает качество ответов. Контур самопроверки: агент оценивает свой ответ по формализованному набору критериев до выдачи — противоречия с ранее сказанным, подмена прямого ответа обтекаемым, деградация связности — и переформулирует сам.

Оркестрация и маршрутизация. Голосовое или текстовое сообщение → классификация → маршрутизация в нужный обработчик → исполнение → отчёт. Переиспользуемая база skill-модулей, из которой агент собирается под задачу без разработки с нуля. Агент авто-онбординга: обходит сайт клиента до 8 страниц, извлекает профиль компании, обогащает данные через внешний справочник и настраивает бота без участия человека.

Учёт эксплуатации LLM — «Хранитель». Собственный контур: сборщики на боевых серверах отдают расход по провайдерам в ingest-эндпоинт, защищённый токеном, панель сводит стоимость по проектам вместе с выручкой и расходами на хостинг. Плюс зашифрованное хранилище ключей — AES-GCM, PBKDF2 с 600 000 итераций, срок до ротации по каждому секрету. Два интерфейса над одной базой: Telegram-бот с голосовым вводом и веб-панель на FastAPI, синхронизация с серверами — systemd-юнитами.

Сейчас в работе: адаптер подписочного доступа для агентной среды OpenClaw — инструментарий уровня фреймворка, а не прикладной бот.

Про фреймворки честно. RAG и агентов собирал без LangChain и LlamaIndex — на ChromaDB и sentence-transformers напрямую. Поэтому понимаю, что фреймворк делает под капотом: чанкинг, retrieval, сборку контекста, оркестрацию инструментов. Освоить обёртку — вопрос дней; обратный путь так не работает.

СТЕК

Python · FastAPI, Pydantic, asyncio, SQLAlchemy, Alembic, Celery, APScheduler, httpx, aiohttp, pytest

LLM · Anthropic SDK, Google GenAI / Vertex AI, OpenAI, Groq, Cerebras, DeepSeek, OpenRouter, Ollama; проектирование промптов и агентных протоколов, собственный слой вызова инструментов, контроль расхода и стоимости токенов

RAG и векторный поиск · ChromaDB, sentence-transformers, multilingual-e5, стратегии чанкинга и перекрытия, инкрементальная переиндексация, метаданные и фильтры

Данные · PostgreSQL (psycopg, asyncpg), Redis, SQLite

Инфраструктура · Linux, Docker, Docker Compose, nginx, systemd, деплой скриптами по SSH, Cloudflare, DigitalOcean, Яндекс.Облако; мониторинг — systemd, логи, алерты в Telegram, Flower, экспорт метрик Prometheus из приложения

Telegram · aiogram, python-telegram-bot, Pyrogram — боты, платежи, вебхуки, голосовой ввод

ML и медиа · Whisper STT, PyTorch, HuggingFace Transformers, Stable Diffusion, Pillow, FFmpeg, numpy, pandas

Frontend · React, Next.js 14, TypeScript, Tailwind; мобильное — React Native, Expo

Автоматизация · Playwright, BeautifulSoup, Selenium

ОПЫТ

НейроАктив — основатель, AI-инженер

2025 — настоящее время · AI-автоматизация для малого и среднего бизнеса · удалённо

- **B2B SaaS CRM для отделов продаж** в проде на отдельном домене с SSL: лиды из Avito и корпоративной почты в единую воронку, прайс-листы с ролевой логикой наценки, аналитика продаж, AI-бот, панель суперадминистратора. Multi-tenant с изоляцией данных по аккаунтам, двухуровневые правила фильтрации почты, денормализация горячих полей под быструю выдачу, генерация счетов в PDF с расчётом НДС. FastAPI + Next.js 14 + PostgreSQL + Redis, Docker Compose, nginx
- **AI-консультант в проде:** 622 модуля Python, 135 000 строк. Диалоговый контур с долговременной памятью, RAG-поиск по корпусу, генерация PDF-отчётов, приём платежей
- **Агенты, забирающие функции целиком:** подготовка ответов на требования и проверки налоговой; ведение юридического контура компании — договоры, претензии, процессуальные документы, сопровождение арбитражных и гражданских дел
- **Аналитический погодный хаб:** агрегация прогнозов из независимых источников, приведение к единой модели, расчёт производных показателей, графики и алерты. Веб и мобильное приложение поверх одного ядра. FastAPI + PostgreSQL + Alembic + Celery, React, React Native + Expo, Firebase push
- **VPN-сервис с приёмом платежей:** управление узлами Xray/WireGuard, автоматическая выдача конфигураций, редакторский контур на LLM с автопостингом
- **Обратная разработка проприетарного формата чертежей КОМПАС-3D:** разбор `.cdw` напрямую на Python без установленного КОМПАСа — ZIP-контейнер, многопоточный zlib, UTF-16. Аудит и дедупликация базы на 2 000+ единиц, пакетная правка штампов
- **Эксплуатация 6+ боевых серверов:** провижининг, systemd-юниты, nginx, деплой по SSH, мониторинг, ротация секретов, разбор инцидентов

Двадцать лет до этого — управление процессами и данными

Tele2, руководитель контакт-центра (2006–2012). Путь от оператора до руководителя 350 подчинённых; обслуживание 28 регионов круглосуточно. Управление строилось на цифрах: анализ показателей, отслеживание просадок в реальном времени, прогнозирование входящего трафика, отчётность с дискретностью 15 минут, real-time визуализация показателей в операторских залах. Отток персонала снижен с 7,5% до 1,2% в месяц, время обслуживания — с 120 до 100 секунд. Отдельно построил с нуля бэк-офис на 50 человек: письменные претензии абонентов 36 регионов, удовлетворённость решением с 65% до 85%.

Ростелеком (2015–2016). Абонентский отток нормализован до целевых показателей за полгода. Централизация архивов из 50+ населённых пунктов.

Собственные B2B-компании (2016 — настоящее время): СтройЮгКомплект, ЭКОРЕСУРС, ПОСЕЙДОН. Все бизнес-процессы с нуля. Оборот до 30 млн ₽ в год, сделок на 150+ млн ₽, база 1 500 активных клиентов. Собственная CRM, онлайн-просчёт заказа в момент звонка, маркетинг и вёрстка своими силами.

ОБРАЗОВАНИЕ

Южный федеральный университет — химик, специальность «Химия», специализация «Коллоидная химия», 2004–2010. Расчётная часть программы: математика 680 часов, **численные методы и программирование**, квантовая механика и квантовая химия, кристаллохимия, строение вещества. Плюс квалификация преподавателя, дополнительное к высшему, 2005–2010, 1 400 часов.

ДГТУ — аспирантура, не окончена. «Управление в стиле коучинг» 2011 · немецкий B1.

Готов к техническому интервью с разбором любого проекта: архитектура, код, деплой, инциденты в проде и как они разбирались.