

18+

ФЕДОР ДОНОВАН

ЛОКАЛЬНЫЙ ИИ В ОФИСЕ



ИИ

Федор Донован

Локальный ИИ в офисе

<https://litres.ru/74465208>

ISBN 9785007121767

Аннотация

В этой книге вы узнаете как установить в организации локальную версию ИИ. Последовательно по шагам подключить к 1С и любым другим базам. И получать любые отчеты за секунды. Например отчет по продажам, остатки и движение товара, прогноз закупок и неликвиды, сверка расчетов с поставщиками, распознавание накладных и счетов, анализ маржинальности по товарам и менеджерам, подготовка писем и ответов клиентам.

Содержание

Глава 1. Почему ваш Excel больше не справляется: точка невозврата	5
Глава 2. Что реально умеет локальный ИИ в торговле: 7 сценариев, которые окупятся за месяц	13
Глава 3. Один сервер, пятнадцать компьютеров: архитектура, которая не зависит от интернета	25
Глава 4. Железо: выбираем сервер под 15 пользователей	32
Конец ознакомительного фрагмента.	38

Локальный ИИ в офисе

Федор Донован

© Федор Донован, 2026

ISBN 978-5-0071-2176-7

Создано в интеллектуальной издательской системе Ridero

Глава 1. Почему ваш Excel больше не справляется: точка невозврата

Пятница, 18:47. Офис торговой компании «Северный путь». За окном уже темнеет. Сергей, коммерческий директор, в третий раз перепроверяет сводную таблицу в Excel. Недельный отчёт по продажам должен уйти собственнику до 19:00, но цифры не сходятся: в 1С одни остатки, в складской программе — другие, а менеджер по закупкам ещё не скинул данные по последним приходам. Телефон разрывается: бухгалтерия просит подтвердить отгрузку, логист уточняет, сколько коробок кофе осталось на складе в Новосибирске, а собственник пишет в мессенджер: «Жду отчёт».

Сергей открывает пять вкладок, копирует ячейки, вручную сводит цифры из трёх источников. На вопрос логиста он отвечает наугад: «Остатков должно хватить, завтра уточню». Все замирают в ожидании. А в соседнем кабинете уже стоит готовый локальный ИИ-сервер — небольшая коробка с видеокартой, подключённая к локальной сети. Если бы Сергей просто спросил у неё: «Сколько у нас осталось позиции X на складе в Новосибирске?» — через четыре секунды получил бы точный ответ, а отчёт был бы сформирован ещё полчаса назад.

Знакомая картина? Скорее всего, да. Большинство малых

и средних торговых компаний живут в режиме ручного сбора данных, бесконечных Excel-файлов и нервных пятничных вечеров. Мы привыкли считать это нормой, но на самом деле это точка невозврата: объём информации давно превысил возможности человеческой обработки. И единственный выход — не нанимать ещё одного аналитика, а дать компьютеру делать то, что он умеет лучше всех: читать, считать, сравнивать и отвечать на вопросы за секунды.

Эта книга — практическое руководство, которое поможет вам построить собственного ИИ-помощника внутри вашей компании. Не в облаке, не по подписке, а на физическом сервере, стоящем в вашем офисе и подключённом к 15 компьютерам. Вы узнаете, как за несколько недель внедрить систему, которая возьмёт на себя рутинные отчёты, анализ продаж, контроль остатков и подготовку документов. Но, прежде чем мы перейдём к железу и настройкам, давайте честно посмотрим на то, что происходит в вашей организации сегодня.

Боль, которую вы чувствуете каждый день

Давайте посчитаем время, которое ваши сотрудники тратят на рутинные операции. Возьмём типичную торговую компанию с 15 сотрудниками: 5 менеджеров по продажам, 2 бухгалтера, 2 логиста, 1 закупщик, 1 маркетолог, 1 руководитель и вспомогательный персонал.

Отчёты. Каждый менеджер ежедневно готовит отчёт по своим клиентам и продажам — минимум 30 минут в день.

Это 2,5 часа в неделю на человека, или 12,5 часов суммарно. Руководитель собирает общий отчёт по продажам за неделю — ещё 2—3 часа, плюс сверка с планом. Бухгалтерия вручную переносит данные из банковских выписок в 1С — 1 час в день. Итого за неделю только на отчёты уходит около 20—25 человеко-часов.

Остатки и склад. Логист и закупщик постоянно отвечают на вопросы: «Есть ли товар?», «Когда приедет?», «Сколько заказать?». Они звонят на склад, смотрят в 1С, сверяются с таблицами. Минимум 30 минут в день на уточнения. Неделя — ещё 2,5 часа.

Работа с документами. Входящие накладные, счета, акты — бухгалтерия вручную вносит их в учётную систему, сверяет с заказами, ищет расхождения. Даже при использовании сканера и распознавания текста процесс занимает 15—20 минут на документ, а их бывает 10—15 в день. Это ещё 2—3 часа ежедневно.

Аналитика. Собственник или руководитель раз в неделю просит отчёт по маржинальности, дебиторке, неликвидам. Менеджер по аналитике (если он есть) тратит 3—4 часа на выгрузку и свод данных. Чаще всего таких специалистов нет, и эту работу делает кто-то «на коленке» в Excel, допуская ошибки.

Сложите всё вместе — получится 40—50 человеко-часов в неделю только на рутинные операции, связанные с данными. При средней зарплате 60 000 рублей в месяц (≈ 350 руб./

час) это около 15 000—17 000 рублей в неделю, или 65 000—70 000 рублей в месяц. В год — порядка 800 000 рублей. И это без учёта ошибок, которые приводят к неверным закупкам, потерянным клиентам и штрафам.

Но деньги — не главное. Главное — время и нервы. Ваши сотрудники выгорают, руководитель не видит реальной картины, решения принимаются на основе устаревших данных. Компания буксует, хотя рынок требует скорости.

Почему облачные ИИ — не выход

Вы наверняка слышали про ChatGPT, YandexGPT, GigaChat и другие облачные нейросети. Они впечатляют, но для бизнеса несут серьёзные риски.

Коммерческая тайна. Отправляя в облачный сервис свои данные о продажах, закупочных ценах, клиентах и поставщиках, вы фактически передаёте их третьей стороне. Даже если провайдер обещает конфиденциальность, юридически вы не контролируете, где хранятся ваши документы, кто имеет к ним доступ и не будут ли они использованы для обучения чужих моделей. Для торговой компании утечка закупочных цен или клиентской базы может стать фатальной.

Зависимость от интернета. Если у вас пропадёт связь, облачный ИИ станет бесполезен. А в торговле отключение интернета даже на час может остановить отгрузки. Локальная система работает всегда — по локальной сети, без внешних соединений.

Регулярные платежи. Подписка на корпоративные версии

облачных ИИ стоит от 500 до 2000 рублей за пользователя в месяц. На 15 сотрудников — от 90 000 до 360 000 рублей в год. При этом вы не владеете ни моделью, ни данными. Локальный сервер окупается за 6—12 месяцев, а дальше работает бесплатно.

Ограниченная кастомизация. Облачные сервисы не знают специфики вашей компании: названия складов, номенклатуру, правила расчёта маржи, формы отчётов. Вам придётся объяснять им всё каждый раз заново, и они всё равно будут ошибаться в деталях. Локальная модель может быть обучена или настроена на ваши внутренние документы и терминологию.

Контроль версий и обновлений. Облачный провайдер может в любой момент изменить модель, интерфейс или условия использования. Вы не можете зафиксировать рабочую версию или откатиться назад. Локальная система полностью под вашим контролем: вы решаете, когда обновлять модель, какие данные ей давать, какие ответы она может генерировать.

Что вы получите, дочитав эту книгу

Мы построим локального ИИ-помощника для торговой организации на 15 компьютеров. Это будет система, которая:

Отвечает на вопросы по данным компании — остатки, продажи, задолженности, цены, движение товара. Вопросы можно задавать на естественном языке: «Сколько у нас осталось кофе Lavazza на складе „Центральный“?», «Какие това-

ры не продавались больше 60 дней?», «Кто из менеджеров выполнил план на этой неделе?»

Готовит отчёты — ежедневные, еженедельные, по запросу. В формате Excel, Word или просто текстом в чате. Отчёт по продажам за месяц будет формироваться за 2—3 минуты, а не за 2 часа.

Анализирует данные — считает маржинальность, оборачиваемость, ABC/XYZ-анализ, прогнозирует спрос и закупки, выявляет неликвиды и кассовые разрывы.

Работает с документами — распознаёт сканы накладных и счетов, сверяет их с данными учётной системы, находит расхождения.

Помогает в коммуникациях — готовит черновики ответов клиентам, коммерческие предложения, письма поставщикам, используя данные из вашей базы.

Всё это будет работать локально, без интернета, с разграничением прав доступа. Бухгалтер не увидит зарплаты менеджеров, а менеджер — закупочные цены. Каждый запрос будет логироваться, и вы всегда сможете проверить, кто и что спрашивал.

Книга построена как пошаговое руководство. В первых трёх главах мы разберём, зачем это нужно, какие задачи можно решить и какова архитектура системы. Затем перейдём к практике:

Глава 4—5 — выбор и настройка оборудования, установка операционной системы и базового ПО.

Глава 6 — запуск первой языковой модели, которая умеет общаться на русском языке.

Глава 7 — подключение данных компании: Excel, CSV, PDF, выгрузки из 1С.

Глава 8 — интеграция с 1С и автоматизация выгрузок.

Глава 9 — настройка ролей и прав доступа.

Глава 10 — готовые шаблоны запросов (промптов) для типовых задач торговли.

Главы 11—13 — практические сценарии: отчёты о продажах, склад и закупки, бухгалтерия и финансы.

Глава 14 — безопасность и резервное копирование.

Глава 15 — план внедрения от пилота до ежедневной работы.

Вам не нужно быть программистом. Все команды, скрипты и конфигурации будут приведены полностью — достаточно копировать и вставлять, следуя пояснениям. В конце каждой главы — контрольный список, чтобы вы могли проверить, что всё сделано правильно.

Почему именно сейчас

Технологии локальных языковых моделей за последние два года сделали колоссальный скачок. Ещё в 2022 году казалось, что собственный ChatGPT — удел крупных корпораций. Сегодня же модель уровня GPT-3.5 можно запустить на видеокарте за 50 000 рублей, и она будет отвечать на русском языке не хуже облачных сервисов. Появились удобные инструменты с веб-интерфейсом: Ollama, Open WebUI,

AnythingLLM. Всё это бесплатно и с открытым исходным кодом.

Оборудование тоже подешевело: сервер с видеокартой на 16 ГБ VRAM обойдётся в 120—150 тысяч рублей — это разовая инвестиция, которая окупится за несколько месяцев. А если у вас уже есть мощный ПК с игровой видеокартой, можно начать вообще без затрат.

Торговые компании, которые внедряют такие системы уже сейчас, получают конкурентное преимущество: они быстрее принимают решения, меньше ошибаются, экономят время сотрудников. Те, кто откладывает, продолжат тонуть в Excel и терять деньги на рутине.

В следующей главе мы разберём семь конкретных сценариев, в которых локальный ИИ приносит максимальную пользу именно торговой организации. Вы увидите реальные примеры запросов и ответов, а также оцените экономию времени для каждого сценария. После этого станет окончательно ясно, что внедрение — не роскошь, а необходимость.

А пока отложите свой недельный отчёт и задайте себе один вопрос: сколько ещё пятничных вечеров вы готовы потратить на сведение цифр из трёх программ? Ответ, скорее всего, — ни одного. Тогда переверните страницу и начнём строить вашего электронного помощника.

Глава 2. Что реально умеет локальный ИИ в торговле: 7 сценариев, которые окупятся за месяц

Это не робот-бухгалтер, который заменит ваших сотрудников. Это неутомимый аналитик, который ночью читает все накладные, выгрузки и таблицы, а утром кладёт вам на стол готовый отчёт. Он не устаёт, не ошибается из-за невнимательности и не уходит в отпуск. Локальный ИИ — это инструмент, который берёт на себя самую нудную часть работы с данными, освобождая людей для задач, требующих человеческого суждения, переговоров и творчества.

В этой главе мы разберём семь конкретных сценариев, в которых локальный ИИ приносит максимальную пользу торговой организации. Для каждого сценария — пример запроса, как это работает, какая экономия времени достигается. Вы увидите, что большинство задач, на которые сегодня уходят часы, решаются за минуты. После прочтения вы сможете выбрать два-три сценария для пилотного внедрения в своей компании.

Сценарий 1. Отчёт по продажам за 2 минуты вместо 2 часов

Проблема. Каждый понедельник руководитель или менеджер собирает отчёт по продажам за прошлую неделю: выгружает данные из 1С, сводит их с планами, добавляет данные по клиентам и товарным группам, форматирует в Excel, отправляет собственнику. Это занимает от 1,5 до 3 часов.

Как решает локальный ИИ. Вы задаёте вопрос в чат-интерфейсе: «Сделай отчёт по продажам за прошлую неделю: общая выручка, выполнение плана по каждому менеджеру, топ-10 товаров по выручке, сравнение с предыдущей неделей». ИИ обращается к заранее подготовленным выгрузкам из 1С (обновляются автоматически каждую ночь), анализирует данные и выдаёт готовый отчёт в текстовом виде, а при необходимости — в формате Excel или Word. Время: от 1 до 3 минут.

Пример запроса и ответа.

Вы: «Подготовь еженедельный отчёт по продажам за период с 10 по 16 марта. Включи: общую выручку, выполнение плана по менеджерам, отклонение от прошлой недели, топ-10 SKU по выручке, продажи по каналам (опт/розница)».

ИИ (через 40 секунд):

«Отчёт по продажам за неделю 10—16 марта.

Общая выручка: 4 820 000 руб. (план — 5 000 000, выполнение 96,4%).

По менеджерам: Иванов — 1 150 000 (92%), Петров — 980 000 (108%), Сидоров — 1 320 000 (101%), Козлов —

780 000 (85%).

Отклонение от прошлой недели: +3,2%.

Топ-10 SKU: кофе Lavazza 1 кг — 320 000 руб., сахар 5 кг — 210 000 руб.,...

Каналы: опт — 3 100 000, розница — 1 720 000.»

Экономия времени: 90—150 минут на каждом отчёте. При еженедельной подготовке — 6—10 часов в месяц только на одного сотрудника.

Сценарий 2. Ответы по остаткам и движению товара

Проблема. Логисты, менеджеры и закупщики постоянно отвечают на вопросы: «Сколько осталось?», «Когда придет?», «Что уже отгружено?». Ответ требует захода в 1С, складскую программу, иногда — звонок на склад. В день таких запросов бывает 10—20, каждый занимает 2—5 минут. Суммарно — до часа в день.

Как решает локальный ИИ. Вы подключаете к ИИ выгрузку остатков и движений товара (обновляется, например, каждый час). Теперь любой сотрудник может спросить в чате: «Остаток кофе Lavazza 1 кг на складе „Центральный“ с учётом зарезервированного?» ИИ мгновенно даёт ответ, может показать историю движения за период.

Пример запроса и ответа.

Менеджер: «Сколько у нас осталось сахара 5 кг на основном складе?»

ИИ: «На складе „Основной“ остаток сахара 5 кг — 340 упаковок. Свободно (без резерва) — 290. За последние 7

дней отгружено 120, доходов не было. Среднедневные продажи — 17 упаковок, текущего остатка хватит примерно на 17 дней.»

Экономия времени: 30—60 минут в день на уточнениях, плюс снижение ошибок из-за устаревшей информации.

Сценарий 3. Прогноз закупок и выявление неликвидов

Проблема. Закупщик вручную анализирует историю продаж, сезонность, остатки и решает, сколько товара заказать. Часто заказывает «на глазок», что приводит к избытку неликвидов или дефициту ходовых позиций. Анализ неликвидов (товары, которые не продаются 60—90 дней) делается редко, потому что это трудоёмко.

Как решает локальный ИИ. ИИ может автоматически рассчитывать точку заказа для каждого SKU на основе средних продаж, страхового запаса, времени доставки. Может формировать список товаров с низкой оборачиваемостью, оценивать замороженные средства.

Пример запроса и ответа.

Закупщик: «Покажи товары, которые не продавались более 60 дней, с остатком более 20 000 руб., и рассчитай сумму замороженных средств по ним.»

ИИ:

«Найдено 14 позиций с продажами 0 за последние 60 дней. Суммарный остаток в закупочных ценах: 540 000 руб.

Топ-3:

Чай зелёный «Ориент» 100 пак. — остаток 450 шт., заку-

почная цена 95 руб., сумма 42 750 руб.

Печенье «К чаю» 250 г — остаток 1200 шт., цена 54 руб., сумма 64 800 руб.

...

Рекомендую рассмотреть распродажу или возврат поставщику.»

Пример запроса 2.

Закупщик: «Сформируй черновик заказа поставщику „Торгснад“ на следующую неделю по кофе и сахару с учётом текущих остатков, прогноза продаж и минимальной партии.»

ИИ:

«Заказ на 24—28 марта:

Кофе Lavazaz 1 кг — заказать 150 шт. (остаток 120, прогноз продаж 180, мин. партия 50).

Сахар 5 кг — заказать 300 шт. (остаток 340, прогноз продаж 260, мин. партия 100, но ожидается сезонный рост).

...»

Экономия времени: закупщик тратит 3—4 часа в неделю на расчёты; ИИ сокращает это до 30 минут, а качество прогноза повышается.

Сценарий 4. Сверка расчётов с поставщиками

Проблема. Бухгалтерия регулярно сверяет взаиморасчёты: сколько поставщик нам поставил, сколько мы оплатили, какие накладные не закрыты. Это сопоставление данных из 1С, банковских выписок, входящих накладных. Ошибки

приводят к переплатам или задолженностям, которые всплывают в конце квартала.

Как решает локальный ИИ. ИИ может автоматически сопоставлять выгрузки из 1С (поступления товаров и платежей) и находить расхождения: неоплаченные накладные, переплаты, отсутствующие документы. Может формировать акт сверки.

Пример запроса и ответа.

Бухгалтер: «Сверь расчёты с поставщиком ООО „Торгснаб“ за февраль. Найди расхождения между накладными и оплатами.»

ИИ:

«По ООО «Торгснаб» за февраль:

— Поставки по накладным: 1 250 000 руб. (12 накладных).

— Оплаты: 1 000 000 руб. (3 платежа).

— Найдено расхождение: накладная №245 от 15.02 на 150 000 руб. не оплачена.

— Переплата по накладной №210 от 05.02: 100 000 руб. (оплачено 200 000 при поставке 100 000).

Итого задолженность перед поставщиком: 150 000 руб. Рекомендую запросить акт сверки.»

Экономия времени: сверка одного поставщика занимает 1—2 часа, при наличии 20—30 поставщиков — это дни работы. ИИ сокращает время до нескольких минут.

Сценарий 5. Распознавание накладных и счетов

Проблема. Входящие бумажные документы (накладные, счета, акты) нужно вручную вносить в учётную систему. Даже при использовании сканера и OCR приходится проверять и исправлять ошибки. Это монотонная работа, на которую уходит значительная часть дня бухгалтера.

Как решает локальный ИИ. Локальная OCR-модель (например, на базе PaddleOCR) распознаёт текст, а языковая модель извлекает нужные поля: дату, номер, поставщика, номенклатуру, количество, цены, сумму. Всё это автоматически подставляется в форму или сравнивается с заказом.

Пример запроса и ответа.

Бухгалтер загружает скан накладной в чат и пишет: «Распознай этот документ и сверь с заказом №123.»

ИИ:

«Накладная №789 от 18.03 от ООО «Торгснад». Позиции:
1) Кофе Lavazza 1 кг — 50 шт., цена 450 руб., сумма 22 500 руб.
2) Сахар 5 кг — 100 шт., цена 120 руб., сумма 12 000 руб.
Итого: 34 500 руб.

Сверка с заказом №123: совпадение по позициям и количеству. Расхождений нет. Документ готов к проводке.»

Экономия времени: вместо 10—15 минут ручного ввода — 1—2 минуты на проверку. При 10 документах в день экономия до 2 часов ежедневно.

Сценарий 6. Анализ маржинальности по товарам и менеджерам

Проблема. Руководство хочет знать, какие товары и какие

менеджеры приносят наибольшую прибыль. Расчёт маржинальности требует учёта закупочных цен, скидок, бонусов, транспортных расходов. Обычно это делается раз в месяц и занимает несколько часов, а текущая картина остаётся неясной.

Как решает локальный ИИ. ИИ может рассчитывать маржинальность по каждой продаже на основе данных о закупочной цене (из закупок) и цене продажи (из отгрузок). Может группировать по товарам, группам, менеджерам, клиентам. Может выявлять нерентабельные позиции.

Пример запроса и ответа.

Руководитель: «Покажи маржинальность по менеджерам за последний месяц, с разбивкой по валовой прибыли и марже в процентах.»

ИИ:

«Маржинальность за период 01.02—28.02:

Иванов: выручка 4 500 000, валовая прибыль 1 350 000, маржа 30%.

Петров: выручка 3 800 000, валовая прибыль 950 000, маржа 25%.

Сидоров: выручка 4 100 000, валовая прибыль 1 640 000, маржа 40%.

...

Обратите внимание: у Петрова низкая маржа из-за большого объёма продаж со скидкой. Рекомендую проанализировать его сделки со скидкой более 15%.»

Экономия времени: формирование такого отчёта вручную занимает 2—4 часа, ИИ делает за 1—2 минуты. Плюс возможность получать анализ чаще, чем раз в месяц.

Сценарий 7. Подготовка писем и ответов клиентам

Проблема. Менеджеры тратят время на написание стандартных писем: подтверждение заказа, счёт на оплату, ответ на вопрос о наличии, коммерческое предложение. Эти письма можно шаблонизировать, но всё равно требуют времени на подстановку данных и формулировку.

Как решает локальный ИИ. ИИ может генерировать тексты писем на основе шаблонов и данных из системы. Например, узнав у ИИ остаток товара, можно сразу попросить: «Напиши клиенту „ООО Ромашка“ письмо о готовности отгрузки 50 упаковок кофе по цене 450 руб., счёт будет направлен отдельно». ИИ сгенерирует вежливое профессиональное письмо.

Пример запроса и ответа.

Менеджер: «Напиши клиенту ООО „Виктория“ ответ на их запрос о наличии кофе Lavazza 1 кг. У нас есть 200 шт., цена 480 руб., доставка завтра.»

ИИ:

«Уважаемый Иван Петрович!

Благодарим за интерес к нашей продукции. Сообщаем, что кофе Lavazza 1 кг есть в наличии в количестве 200 упаковок. Цена за упаковку — 480 рублей. Возможна отгрузка уже завтра.

Если вас устраивают условия, просим подтвердить заказ, и мы выставим счёт.

С уважением,

Менеджер отдела продаж

ООО «Северный путь»»

Экономия времени: каждое письмо — 5—10 минут, при 10 письмах в день — до 1,5 часов. ИИ сокращает до 30 секунд на генерацию и 1 минуты на проверку.

Границы возможного: что ИИ делать не должен

Несмотря на всю полезность, локальный ИИ не всесильен и не должен заменять человеческое суждение в критически важных областях. Чётко определите границы его применения.

1. Принятие юридически значимых решений. ИИ может подготовить проект договора, сформировать акт сверки, но подписывать документы, принимать решения о списании долгов или претензионной работе должен человек. Все финансовые и юридические действия требуют контроля и ответственности.

2. Прямое подключение к банку и проведение платежей. ИИ может анализировать выписки, прогнозировать кассовые разрывы, но не должен иметь права создавать платёжные поручения и отправлять их в банк без участия бухгалтера. Безопасность — превыше всего.

3. Замена бухгалтера или финансового директора. ИИ не может нести ответственность за правильность учёта, не мо-

жет интерпретировать нестандартные ситуации (например, возврат брака, взаимозачёт). Его выводы — это информация для принятия решения, а не само решение.

4. Гарантированно точные прогнозы. ИИ строит прогнозы на основе исторических данных, но не учитывает внезапные изменения рынка, действия конкурентов, форс-мажоры. Прогнозы нужно воспринимать как ориентир, а не как истину.

5. Работа с неструктурированными или противоречивыми данными без контроля. Если в 1С хаос, документы не проведены, остатки неверны, ИИ не сможет исправить это волшебным образом. Сначала нужно навести порядок в учёте, а ИИ лишь автоматизирует работу с уже корректными данными.

6. Замена живого общения с клиентами и партнёрами. ИИ может подготовить текст, но вести переговоры, разрешать конфликты, выстраивать отношения должен человек.

Какой практический результат вы получаете

К концу этой главы вы должны были убедиться, что локальный ИИ — это не абстрактная технология, а конкретный инструмент с измеримой экономией времени и денег. Семь сценариев покрывают основные рутинные задачи торговой организации: отчёты, остатки, закупки, сверки, документы, аналитика, коммуникации.

Теперь ваша задача — выбрать 2—3 сценария, которые наиболее актуальны для вашей компании. Не нужно пытаться

ся внедрить всё сразу. Начните с того, что болит сильнее всего: например, еженедельные отчёты и ответы по остаткам. Именно на этих задачах вы быстрее всего увидите эффект и получите поддержку коллектива.

В следующей главе мы разберём, как устроена система технически: один сервер, пятнадцать компьютеров, архитектура, которая не зависит от интернета. Вы поймёте, что для старта не нужен суперкомпьютер и команда программистов — достаточно вашего желания и нескольких вечеров на настройку.

А пока запишите на листке бумаги три задачи, которые отнимают у вас больше всего времени. Скорее всего, именно они станут вашими пилотными сценариями.

Глава 3. Один сервер, пятнадцать компьютеров: архитектура, которая не зависит от интернета

Вы думаете, для запуска собственного ИИ нужен супер-компьютер за полмиллиона рублей, команда программистов и серверная комната с кондиционером? Отбросьте этот стереотип. Скорее всего, вам хватит обычного сервера за 120—150 тысяч рублей, а иногда — даже старого офисного ПК с добавлением игровой видеокарты. В этой главе мы разберём, как устроена система изнутри, чтобы вы убедились: внедрение локального ИИ в компанию на 15 компьютеров — задача реальная, выполняемая за несколько вечеров и не требующая глубоких технических знаний.

Простая схема: сервер, сеть и браузер

Представьте себе обычную локальную сеть вашего офиса. Компьютеры сотрудников соединены между собой через роутер или коммутатор. К этой же сети мы добавляем ещё один компьютер — сервер, на котором будет работать ИИ. Всё. Никаких дополнительных проводов, облачных сервисов или сложного оборудования.

Сервер — это обычный системный блок, но с одной важной деталью: мощной видеокартой (GPU). Именно видеокарта выполняет все вычисления нейросети. Для работы

языковой модели уровня GPT-3.5 достаточно видеокарты с 16—24 гигабайтами видеопамяти. Такие карты стоят от 40 до 150 тысяч рублей. В главе 4 мы подробно рассмотрим варианты, а пока просто запомните: это не сервер за миллионы, а скорее мощный игровой компьютер.

Сотрудники заходят в чат-интерфейс, как в Telegram. На каждом рабочем компьютере уже есть браузер — Chrome, Firefox, Edge. Никакого дополнительного программного обеспечения устанавливать не нужно. Сотрудник открывает браузер, вводит локальный адрес сервера (например, <http://192.168.1.10:3000>) и попадает на страницу с чатом, очень похожую на ChatGPT или Telegram. Там он пишет свой вопрос на обычном русском языке и через несколько секунд получает ответ. Всё происходит внутри вашей локальной сети, как если бы вы открыли внутренний сайт компании.

Один сервер — пятнадцать пользователей. Современные языковые модели умеют обрабатывать запросы по очереди или параллельно. При использовании видеокарты с 16 ГБ VRAM и модели среднего размера (7—14 миллиардов параметров) сервер легко выдерживает одновременную работу 5—10 пользователей без заметных задержек. Если все 15 сотрудников обращаются к ИИ не чаще, чем раз в несколько минут, нагрузка будет комфортной. Для пилотного запуска этого достаточно.

Принцип работы: модель, документы и роли

Как же сервер отвечает на вопросы о ваших данных?

Внутри работают три ключевых компонента.

1. Языковая модель. Это «мозг» системы. Она умеет понимать русский язык, рассуждать, писать тексты и анализировать информацию. Мы будем использовать открытые модели, которые можно свободно скачать и запустить: Qwen, Llama, Mistral, Saiga и другие. Они не уступают облачным аналогам в базовых задачах, а для специфических бизнес-задач их можно дообучить на ваших документах. Модель работает локально, на вашем сервере, и не отправляет данные никуда.

2. Векторная база данных и индексация документов. Языковая модель не знает содержимого ваших Excel-файлов и выгрузок из 1С. Чтобы она могла отвечать на вопросы по вашим данным, мы используем технологию RAG (Retrieval-Augmented Generation — генерация с дополнением из поиска). Все ваши документы (Excel, CSV, PDF, Word, выгрузки из 1С) проходят процесс индексации: текст из них разбивается на фрагменты, каждый фрагмент преобразуется в числовой вектор (embedding) и сохраняется в специальной векторной базе данных. Когда пользователь задаёт вопрос, система ищет в векторной базе самые похожие фрагменты документов, подставляет их в запрос к языковой модели и просит сформулировать ответ на основе только этой информации. Благодаря этому ИИ отвечает по вашим актуальным данным, а не выдумывает.

3. Ролевой доступ. У каждого сотрудника есть своя учёт-

ная запись с логином и паролем. Администратор определяет, какие документы и функции доступны каждой роли. Например, менеджеры по продажам видят остатки и цены продажи, но не видят закупочные цены и зарплаты. Бухгалтер видит финансовые документы, но не видит переписку с клиентами. Это реализуется через настройку прав в веб-интерфейсе и разграничение доступа к векторным индексам. Все запросы логируются, поэтому вы всегда можете проверить, кто и что спрашивал.

Сравнение с облаком: почему локально выгоднее

Давайте честно сравним локальную систему с облачными сервисами, такими как ChatGPT, YandexGPT или GigaChat, по четырём критериям.

Стоимость. Облачные корпоративные тарифы стоят от 500 до 2000 рублей за пользователя в месяц. Для 15 сотрудников это от 90 000 до 360 000 рублей в год. Локальный сервер — разовая инвестиция 120—150 тысяч рублей. Он окупается за первый год, а дальше работает бесплатно. Электроэнергия: видеокарта потребляет 200—300 Вт, при круглосуточной работе это примерно 500—700 рублей в месяц. Итого годовые затраты после покупки — менее 10 000 рублей на электричество.

Скорость. Облачный сервис зависит от скорости вашего интернета и загруженности серверов провайдера. Локальная система отвечает за 1—5 секунд, потому что данные не покидают офис и нет задержек сети. Даже при медленном ин-

тернете или его полном отсутствии скорость остаётся высокой.

Приватность. Отправляя данные в облако, вы передаёте коммерческую информацию третьей стороне. Вы не контролируете, где хранятся ваши документы, кто имеет к ним доступ, не используются ли они для обучения чужих моделей. Локальная система гарантирует: все данные остаются внутри вашей сети. Вы можете провести аудит, убедиться, что никакая информация не уходит наружу, и при необходимости полностью отключить сервер от интернета.

Контроль. В облаке вы зависите от политики провайдера: он может изменить модель, интерфейс, условия. В локальной системе вы сами решаете, какую модель использовать, когда обновлять, какие данные подключать. Вы можете дообучить модель на своих документах, настроить её под свою терминологию, добавить собственные шаблоны отчётов.

Возможность полностью отключить интернет

Одно из главных преимуществ локальной архитектуры — независимость от внешнего мира. После того как система установлена и все необходимые модели и библиотеки скачаны, вы можете физически отключить сервер от интернета. ИИ продолжит работать, потому что все вычисления происходят внутри вашей локальной сети. Сотрудники так же будут заходить в чат по внутреннему адресу и получать ответы.

Зачем это нужно? Во-первых, безопасность: сервер, не подключённый к интернету, невозможно взломать удалён-

но, с него невозможно украсть данные. Во-вторых, стабильность: ваша работа не зависит от провайдера, сбоев связи или блокировок. В-третьих, соответствие требованиям: для некоторых отраслей (например, работа с персональными данными) критически важно, чтобы информация не покидала периметр компании.

Конечно, для удобства администрирования можно оставить серверу доступ в интернет для получения обновлений, но это не обязательно. В главе 14 мы подробно обсудим вопросы безопасности и способы изоляции.

Практический результат

Итак, вы теперь понимаете, что архитектура локального ИИ для 15 компьютеров проста и доступна:

Один сервер с GPU — центральный узел, выполняющий все вычисления.

Локальная сеть — уже есть в вашем офисе.

Браузер на каждом компьютере — интерфейс для сотрудников, ничего устанавливать не нужно.

Открытые программные компоненты — модель, векторная база, веб-интерфейс — всё бесплатно и свободно распространяется.

Вам не нужно быть программистом, чтобы это внедрить. Все шаги мы разберём детально, с командами и скриншотами. В следующей главе мы перейдём к выбору оборудования: вы узнаете, какую видеокарту купить, сколько нужно оперативной памяти и какой процессор выбрать, чтобы система

работала быстро и надёжно. Вы сможете составить конфигурацию под свой бюджет — от минимальной до продвинутой.

А пока подумайте, есть ли у вас в организации старый, но мощный компьютер, который можно превратить в ИИ-сервер. Возможно, ваш путь к собственному интеллектуальному помощнику начнётся уже сегодня.

Глава 4. Железо: выбираем сервер под 15 пользователей

Вы решились. Вы готовы выделить угол в офисе, бюджет и несколько вечеров на настройку. Первый практический шаг — собрать или выбрать сервер, который будет тянуть локальный ИИ для пятнадцати сотрудников. В этой главе мы разберём каждый компонент, поговорим о требованиях и соберём три конфигурации: минимальную, оптимальную и с запасом. Никакой воды — только то, что действительно важно.

Почему именно видеокарта — сердце системы

Вся работа языковой модели — это миллиарды матричных операций, которые выполняются на графическом процессоре (GPU). Центральный процессор (CPU) тоже участвует: он обрабатывает запросы к векторной базе, управляет веб-интерфейсом, но основная вычислительная нагрузка лежит на видеокарте. Поэтому ключевой параметр — объём видеопамати (VRAM). Именно он определяет, какую модель вы сможете запустить.

Для качественной работы на русском языке с приемлемой скоростью нам нужна модель с 7—14 миллиардами параметров. Такие модели в квантованном виде (сжатые без сильной потери качества) занимают от 5 до 12 ГБ видеопамати. Плюс запас под контекст (историю диалога) и параллельные запро-

сы. Отсюда правило: видеокарта должна иметь минимум 16 ГБ VRAM, а лучше 24 ГБ. Именно поэтому мы рассматриваем карты именно с такими объёмами.

Остальные компоненты сервера подбираются так, чтобы не создавать узких мест и работать стабильно 24/7.

Компоненты сервера: что и зачем

Процессор (CPU)

Для обслуживания 15 пользователей не нужен топовый процессор. Основная работа уходит на GPU, а CPU занимается:

- обработкой HTTP-запросов от сотрудников;

- поиском по векторной базе (это тоже вычисления, но не такие тяжёлые);

- запуском дополнительных скриптов (выгрузки из 1С, планировщик).

Вам подойдёт любой современный 6—8-ядерный процессор, например Intel Core i5—12400/13400 или AMD Ryzen 5 5600/7600. Можно взять и серверный Xeon, но это не обязательно. Главное — поддержка PCIe 4.0 или хотя бы 3.0 для видеокарты.

Вывод: не переплачивайте за CPU. Достаточно среднего уровня.

Оперативная память (RAM)

Языковая модель загружается в VRAM видеокарты, но сама система и приложения требуют оперативной памяти. Для наших задач необходимо минимум 32 ГБ. Если вы пла-

нируете держать в оперативной памяти несколько моделей или большие векторные индексы, берите 64 ГБ. Учтите, что для Docker-контейнеров, веб-интерфейса и самой ОС нужно около 8—10 ГБ.

Вывод: 32 ГБ — минимально, 64 ГБ — комфортно.

Видеокарта (GPU)

Это самый дорогой компонент. Требования: минимум 16 ГБ VRAM, желательно 24 ГБ. Чем больше видеопамати, тем более крупную модель вы сможете запустить и тем больше одновременных запросов обработать.

Мы рассмотрим три реальных варианта:

NVIDIA RTX 4060 Ti 16GB — современная игровая карта с 16 ГБ. Отлично подходит для моделей 7—13B в квантованном виде. Низкое энергопотребление (160 Вт), не требует мощного блока питания, доступна в продаже. Цена новой — от 45 000 до 55 000 рублей. Идеальна для начала.

NVIDIA RTX 4090 24GB — топовая карта с 24 ГБ. Тянет модели до 30B, очень быстрая. Но дорогая (от 180 000 до 250 000 рублей), энергопотребление до 450 Вт, требует мощный блок питания и хорошее охлаждение. Имеет смысл, если вы планируете использовать большие модели (30B+) или много одновременных пользователей.

NVIDIA Tesla P40 24GB (б/у) — серверная карта с 24 ГБ, специально для вычислений. Её преимущество — огромный объём VRAM за небольшие деньги (на вторичном рынке от 35 000 до 60 000 рублей). Но есть нюансы: нет видеовыхо-

да (карта не выводит изображение), требует принудительного охлаждения (у неё пассивный радиатор, нужен мощный вентилятор или переходник), поддерживает только PCIe 3.0. Работает с моделями до 13—30В. Отличный бюджетный вариант, если вы готовы повозиться с охлаждением.

Также можно рассмотреть RTX 3090 24GB — бывшая флагманская игровая карта, сейчас на вторичке стоит 70—100 тысяч рублей. Хорошая альтернатива 4090, если бюджет ограничен, а хочется 24 ГБ.

Важно: все карты должны быть от NVIDIA, так как экосистема для инференса (CUDA, TensorRT) развита именно для них. AMD и Intel пока сильно отстают.

Накопители: NVMe SSD

Для быстрой загрузки моделей и работы системы нужен NVMe SSD объёмом минимум 1 ТБ. Скорость чтения — критична, потому что модель загружается в VRAM при старте. Возьмите SSD с интерфейсом PCIe 3.0 или 4.0. Хорошие варианты: Samsung 970/980 Pro, WD Black SN770/SN850. Цена 1 ТБ — от 7 000 до 12 000 рублей.

Документы компании (Excel, PDF, загрузки) можно хранить на обычном HDD объёмом 2—4 ТБ, чтобы не занимать место на быстром накопителе. Но для векторной базы лучше тоже использовать SSD.

Блок питания

Сервер работает круглосуточно, поэтому блок питания должен быть качественным, с запасом по мощности. Рассчи-

тываем: видеокарта 150—450 Вт + остальная система 150—200 Вт. Для RTX 4060 Ti хватит блока на 550—650 Вт, для RTX 4090 — на 850—1000 Вт. Берите блоки известных брендов (Seasonic, Corsair, be quiet!) с сертификатом 80 Plus Gold или Platinum. Это обеспечит стабильность и меньший нагрев.

Охлаждение

Для сервера, работающего 24/7, важно продумать вентиляцию. Корпус должен быть с хорошим потоком воздуха: минимум два вентилятора на вдув и один на выдув. Процессорный кулер — стандартный башенный (например, DeepCool AK400). Для видеокарты достаточно её собственной системы охлаждения, но если у вас Tesla P40, потребуется отдельный мощный вентилятор, направленный на радиатор, или переходник для установки активного кулера. В серверных корпусах часто используются высокооборотистые вентиляторы, но для офиса лучше выбрать тихие модели, чтобы не мешать сотрудникам.

ИБП (источник бесперебойного питания)

Отключение электричества может привести к потере данных и повреждению индексов. Поэтому обязателен ИБП с мощностью, достаточной для корректного завершения работы. Для системы с RTX 4060 Ti (потребление около 300 Вт) подойдёт ИБП на 600—800 ВА. Для RTX 4090 — на 1000—1500 ВА. ИБП должен поддерживать автономную работу хотя бы 5—10 минут, чтобы сервер успел сохраниться и вы-

ключиться. Многие ИБП (например, APC) имеют USB-подключение для автоматического безопасного завершения работы.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.