



Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

ИИ-ассистент для бухгалтера: ChatGPT без утечки данных

Три контура: локальный Ollama + Open WebUI, GigaChat B2B, OpenAI API с ZDR — плюс регламент

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

Суть проблемы

Сотрудники бухгалтерии уже пользуются ChatGPT и GigaChat стихийно — из браузера, без регламента и понимания, куда уходят запросы. В облако улетают ИНН, паспортные данные, договоры и зарплаты — это персональные данные по 152-ФЗ, коммерческая и налоговая тайна. Мы строим управляемый контур: локальная LLM для чувствительного, облачные тарифы с запретом дообучения для рутины, пись...

Почему это важно бизнесу

- Штрафы по 152-ФЗ: до 700 тыс. руб. за обработку ПДн без согласия (ч.2 ст.13.11 КоАП), до 15 млн при утечке — один вставленный в чат договор созд...
- Теневое использование: ИИ уже применяют без ведома руководителя — риск существует независимо от того, «внедряли» вы ИИ или нет
- Договоры и суммы сделок в логах чужого облака — удар по коммерческой тайне и переговорным позициям
- Правильно внедрённый ассистент экономит бухгалтеру 3–5 часов в неделю на письмах, требованиях ФНС и формулах — это прямые деньги

Ключевые параметры реализации

30 дней

максимум хранения запросов в OpenAI API (без дообучения); для регулируемых отраслей — ZDR

по докам OpenAI Platform

30 мин

срок жизни access-токена GigaChat API — закладываем авто-обновление OAuth в прослойке

по докам developers.sber.ru

127.0.0.1

OLLAMA_HOST: API Ollama слушает только loopback, наружу — через reverse-proxy с аутентификацией

Ollama docs + наш стандарт

Q4_K_M

квантование 7-8B моделей: ~4,7 ГБ RAM вместо 15, работает на офисном сервере без GPU

наш стандарт

8192

num_ctx для бухгалтерских задач: дефолтного контекста не хватает на длинное требование ФНС

наш стандарт

pending

DEFAULT_USER_ROLE в Open WebUI: новые учётки не работают, пока их не активирует админ

Open WebUI env-configuration



Локальный контур: Ollama + Open WebUI в серверной клиента

Что настраиваем

Сервер клиента (Windows Server 2022/Debian 12, от 16 ГБ RAM, без GPU) — автономный ассистент без выхода в интернет

Как мы это делаем

- 1 Ставим Ollama, тянем модель: `ollama pull qwen2.5:7b-instruct-q4_K_M` (~4,7 ГБ); замеряем генерацию на CPU — закладываем не менее 8–10 токенов/с
- 2 Фиксируем службу: `OLLAMA_HOST=127.0.0.1:11434`, `OLLAMA_KEEP_ALIVE=30m`, `OLLAMA_NUM_PARALLEL=2` — модель не выгружается между запросами бухгалтерии
- 3 Поднимаем Open WebUI в Docker: первый вход = админ, затем `ENABLE_SIGNUP=false`, `DEFAULT_USER_ROLE=pending`, персональные учётки сотрудников
- 4 В параметрах модели поднимаем `num_ctx` до 8192 — иначе длинные требования ФНС и договоры молча обрезаются
- 5 Интерфейс публикуем только в LAN через nginx с TLS (внутренний CA); порт 11434 наружу не смотрит, фаервол — deny по умолчанию

РЕЗУЛЬТАТ

Запросы физически не покидают офис: ни логов у провайдера, ни вопросов по 152-ФЗ — в такой чат можно вставлять реальные договоры и ОСВ. Качество 7–8B моделей уровня GPT-3.5 закрывает рутину: письма, пояснения, формулы. Внедрение — один рабочий день на готовом сервере.

КЛЮЧЕВОЙ НЮАНС

В доках Ollama (FAQ/envconfig) важен дефолт `keep_alive = 5` минут: после каждого перерыва модель грузится заново. Для офиса ставим 30m, а `OLLAMA_NUM_PARALLEL` считаем от свободной RAM, а не от числа пользователей...



Облачный контур: OpenAI API и GigaChat B2B через прослойку

Что настраиваем

Задачи без чувствительных данных: шаблоны писем, нормы НК, формулы — через корпоративный бот, а не браузерный чат

Как мы это делаем

- 1 OpenAI — только через API: данные не идут в обучение по умолчанию, хранение до 30 дней; для медицины и финансов запрашиваем ZDR-эндпоинты
- 2 API-ключ — в отдельном проекте с месячным лимитом расходов и ротацией; ChatGPT Team/Enterprise: обучение на данных workspace отключено контрактно
- 3 GigaChat: OAuth на `ngw.devices.sberbank.ru:9443/api/v2/oauth`, scope `GIGACHAT_API_B2B`, RqUID на каждый запрос; access-токен живёт 30 минут — рефреш делает прослойка
- 4 Перед отправкой прослойка маскирует реквизиты regex'ами: ИНН (10/12 цифр), расчётные счета (20 цифр), паспорта — подставляет плейсхолдеры

РЕЗУЛЬТАТ

Сотрудники получают один согласованный вход — корпоративный бот вместо десятка личных аккаунтов. У нас — журнал запросов, контроль расходов и договорные гарантии провайдера вместо пользовательского соглашения бесплатного чата.

КЛЮЧЕВОЙ НЮАНС

Ключевая строка в доках OpenAI: данные API не используются для обучения по умолчанию. В GigaChat это гарантирует только платный scope B2B/CORP — freemium-тариф физлица может использовать запросы для улучшения...



Организационный контур: регламент и периметр

Что настраиваем

Вся компания: регламент на одну страницу, блокировка несогласованных сервисов на шлюзе, инструктаж

Как мы это делаем

- 1 Пишем регламент: белый список инструментов, запрещённые данные (ИНН физлиц, паспорта, суммы сделок, договоры, зарплаты, ключи ЭП), ответственное лицо
- 2 На периметре (pfSense/Keenetic) закрываем DNS-имена несогласованных чатов для всех, кроме разрешённых групп — теневое использование гаснет
- 3 Проводим 30-минутный инструктаж на живых примерах: безопасный запрос-шаблон против вставленного договора с реквизитами
- 4 Заводим процедуру инцидента: фиксация, запрос удаления истории у провайдера, юридическая оценка обязанности уведомить Роскомнадзор

РЕЗУЛЬТАТ

«Тихие» пользователи выходят из тени: у людей появляется разрешённый инструмент и понятная граница допустимого, у руководителя — управляемый риск вместо неизвестности. Регламент занимает одну страницу и внедряется за неделю вместе с техническим контуром.

КЛЮЧЕВОЙ НЮАНС

Запрет без альтернативы не работает: пока не выдан удобный корпоративный ассистент, сотрудники вернутся к личным аккаунтам с телефона. Поэтому блокировку на шлюзе включаем только после запуска локального или о...



Подводные камни

✗ Бесплатный чат = данные в обучение

В браузерном ChatGPT и freemium GigaChat запросы по умолчанию могут идти на дообучение; лечится только тарифом с контрактным запретом или работой чер...

✗ Ollama на 0.0.0.0 без аутентификации

У API Ollama нет встроенной auth: открытый 11434 отдаёт модель всей сети. Держим 127.0.0.1, наружу — только reverse-proxy с TLS и логином

✗ Открытая регистрация в Open WebUI

Первый зарегистрировавшийся становится админом. Сразу после создания админа ставим `ENABLE_SIGNUP=false` и `DEFAULT_USER_ROLE=pending`

✗ Дефолтный контекст режет документы

`num_ctx` по умолчанию мал: длинное требование ФНС обрезается молча, модель «не видит» половину текста. Поднимаем до 8192 и следим за расходом RAM

✗ 7B-модель на слабом железе

С 8 ГБ RAM модель уходит в своп — минуты на ответ, сотрудники бросают инструмент. Минимум 16 ГБ, квантование Q4_K_M, замер токенов/с до сдачи

✗ Ключи ЭП и пин-коды в переписке с ИИ

Файлы контейнеров КриптоПро и пин-коды токенов не передаются никуда и никогда — отдельный пункт регламента, проговариваем на инструктаже

✗ API-ключ без лимитов и ротации

Утёкший ключ OpenAI — чужие расходы на вашем счёте. Отдельный проект, месячный лимит, ротация, хранение в секрет-хранилище, а не в коде бота

✗ Маскирование «руками» не работает

Просить сотрудников самим менять реквизиты на заглушки бесполезно — забывают. Маскирование делаем автоматом в прослойке до отправки в облако



Как правильно

МИНИМУМ

- Регламент на 1 страницу: белый список ИИ и перечень запрещённых данных
- Запрет браузерных бесплатных чатов для работы с реальными данными
- Инструктаж 30 минут с примерами безопасных и опасных запросов

НОРМАЛЬНО

- Локальная LLM: Ollama + Open WebUI, модель 7-8B Q4_K_M, сервер от 16 ГБ RAM
- Облако — только через API или Team-тарифы с контрактным запретом дообучения
- ENABLE_SIGNUP=false, DEFAULT_USER_ROLE=pending, персональные учётки
- Блокировка несогласованных ИИ-сервисов на периметровом шлюзе

ХОРОШО

- Прослойка с автоматическим маскированием ИНН, счетов и ФИО до отправки в облако
- ZDR-эндпоинты OpenAI или GigaChat B2B/CORP для регулируемых данных
- Журналирование запросов и ежеквартальный разбор инцидентов
- RAG по внутренней базе (регламенты, шаблоны писем) на локальной модели

Чек-лист самопроверки

☐ Известно ли руководителю, кто из сотрудников уже пользуется ИИ и в каких задачах?

☐ Есть ли письменный регламент с белым списком инструментов и перечнем запрещённых данных?

☐ Закрыт ли API Ollama на loopback (127.0.0.1:11434), а веб-интерфейс — персональными учётками?

☐ Отключена ли саморегистрация в Open WebUI после создания администратора?

☐ Идут ли облачные запросы только через тарифы с контрактным запретом дообучения?

☐ Маскируются ли ИНН, счета и ФИО автоматически до отправки запроса во внешнее облако?

☐ Хватает ли контекста модели (num_ctx 8192+), чтобы документ читался целиком, а не обрезком?

☐ Знает ли каждый сотрудник, что ключи ЭП и пин-коды не передаются в ИИ ни при каких условиях?

☐ Описана ли процедура на случай, когда чувствительные данные уже ушли в чат?

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.



Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Разворачиваем локальный ИИ-контур (Ollama + Open WebUI) на вашем сервере за один рабочий день
- Подключаем GigaChat B2B / OpenAI API и собираем бота-прослойку с маскированием реквизитов
- Пишем регламент использования ИИ и проводим инструктаж сотрудников на живых примерах
- Закрываем периметр: блокировка теневых ИИ-сервисов на шлюзе, доступ только к согласованным
- Сопровождаем: обновление моделей, контроль расходов API, разбор инцидентов

15+

лет в ИТ-поддержке

50

рабочих мест — наш профиль

МТС

дата-центр, Москва



КОНТАКТЫ

Обсудить вашу задачу

Сайт **itfresh.ru**

Телефон **+7 903 729-62-41**

Telegram **@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.



itfresh.ru



Техническая база

- 01** Ollama — FAQ и конфигурация сервера (env-переменные) (docs.ollama.com — 2026)
- 02** Open WebUI — Environment Variable Configuration (docs.openwebui.com — v0.6)
- 03** OpenAI Platform — Data controls (Your data) (platform.openai.com — 2026)
- 04** OpenAI — Enterprise privacy (openai.com — 2026)
- 05** GigaChat API — REST-справочник и авторизация OAuth (developers.sber.ru — 2026)
- 06** Шаблон регламента использования ИИ ITfresh (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.

