ИИ, который отвечает по базе знаний компании

Замер: справится ли модель, работающая на своём железе, с ответами по базе знаний — и насколько она отстаёт от облачной. Проверено на двух базах и 70 вопросах.

16 сентября 2026 Стенд: домашний ПК с RTX 4060 Бизнес Интегратор

59/60
верных ответов у лучшей модели, поместившейся в видеокарту стенда

60/60
у облачного Claude Sonnet 5 для сравнения

0/10
выдумок там, где ответа в базе нет

Разрыв между своим железом и облаком — один вопрос из шестидесяти. Причина этого промаха не в модели, а в поиске по базе: нужный абзац не попал в подборку, которую модель получила.

Зачем это проверяли

К нам приходят с запросом: «пусть бот отвечает клиентам и сотрудникам по нашей базе знаний». Обычный путь — облачная модель, Claude или GigaChat. Но часть компаний не готова отдавать содержимое обращений и внутренних инструкций наружу: служба безопасности, требования к персональным данным, внутренний регламент.

Тогда возникает вопрос, на который до замера мы отвечали предположениями: **насколько хуже будет отвечать модель, работающая на своём сервере?** Этот отчёт даёт цифры вместо предположений.

Как это работает

Вся база знаний не отправляется в модель при каждом вопросе — это было бы дорого и не влезло бы в её память. Работает связка из двух частей: поиск и модель.

- **Один раз при загрузке.** Статьи режутся на фрагменты по 300–900 знаков. Для каждого считается «отпечаток смысла» — набор чисел. Фрагменты и отпечатки лежат в нашей базе данных.

- **На каждый вопрос.** Считается отпечаток вопроса, из базы достаются 4 ближайших по смыслу фрагмента. Только они и вопрос уходят в модель. Это примерно 1500 знаков вместо всей базы.
- **Если ничего не нашлось** — модель не вызывается вообще: отвечает код, что ответа в базе нет, и зовёт оператора.

ЛОКАЛЬНАЯ СХЕМА

1. Вопрос приходит на сервер компании
2. Поиск по базе — там же
3. Модель отвечает на видеокарте этого же сервера
4. Ответ возвращается клиенту

Где данные: ни вопрос, ни база, ни ответ не покидают периметр компании. Интернет нужен только пользователю.
Что нужно: сервер с видеокартой — на нём и сервис с базой, и модель.

ОБЛАЧНАЯ СХЕМА

1. Вопрос приходит на сервер компании
2. Поиск по базе — там же
3. Вопрос и 4 фрагмента уходят к поставщику ИИ
4. Ответ возвращается клиенту

Где данные: наружу уходит вопрос клиента и куски статей. Вся база целиком — нет. Поставщик не дообучается на этих данных, но запрос в его инфраструктуре побывал.
Что нужно: тоже свой сервер — для сервиса, базы и поиска. Но обычный, без видеокарты.

Что стояло на стенде

Обычный домашний компьютер, не сервер: Ryzen 7 5700X3D, 32 ГБ памяти, видеокарта RTX 4060 с 8 ГБ. Модели работают под Linux внутри Windows (WSL2), обслуживает их Ollama.

МОДЕЛЬ	КТО СДЕЛАЛ	УСТРОЙСТВО	ДЛЯ ЧЕГО ОНА
qwen3.5:9b	Alibaba	9 млрд · 6,6 ГБ	Универсальная модель: 201 язык, код, вызов инструментов. Целиком помещается в видеопамять, поэтому самая быстрая

МОДЕЛЬ	КТО СДЕЛАЛ	УСТРОЙСТВО	ДЛЯ ЧЕГО ОНА
gemma4:12b	Google DeepMind	12 млрд · 7,2 ГБ	Текст и картинки, сжата с учётом квантования, поэтому от сжатия до 4 бит теряет меньше. В 8 ГБ влезает не полностью
gemma4:26b-a4b	Google DeepMind	26 млрд · 18 ГБ	Смесь экспертов: из 128 экспертов на каждом шаге работают немногие, около 4 млрд параметров. Качество крупной модели при скорости небольшой
bge-m3	BAAI	0,6 млрд · 1,2 ГБ	Не отвечает, а ищет: превращает текст в отпечатки смысла. Больше 100 языков, включая русский
Claude Sonnet 5 ОБЛАКО	Anthropic	—	Взята как точка отсчёта: сильная облачная модель, которую мы используем в своих сервисах

Как замеряли

Две базы знаний, чтобы увидеть разницу между маленькой аккуратной и большой разнородной.

БАЗА	СТАТЕЙ	ФРАГМЕНТОВ	ЧТО ЭТО
Внутренняя база компании	8	122	Наши инструкции для сотрудников: доступ к рабочим сервисам, регламент работы с кодом, описания сервисов
Хелпдеск Битрикс24	150	1746	Публичные статьи поддержки: телефония, открытые линии, CRM, коробочная версия

По каждой базе — 35 вопросов, заданных как их задал бы живой человек: 30 с ответом в статьях и **5 таких, ответа на которые в базе нет**. Последние важнее остальных: они показывают, начнёт ли модель выдумывать.

- Все модели получали одни и те же найденные фрагменты и одну и ту же инструкцию — сравнивались модели, а не поиск.
- Ответ считался верным, если в нём есть нужный факт и нет лишнего. Все 280 ответов прочитаны человеком: автоматическая проверка по ключевым словам ошибается в обе стороны.
- Замер сделан по одному прогону на модель, без повторов.

Результаты

МОДЕЛЬ	ВНУТРЕННЯЯ БАЗА	ХЕЛПДЕСК	ВСЕГО ВЕРНЫХ ИЗ 60	СЕК / ОТВЕТ
qwen3.5:9b	29/30	26/30	55	1,5
gemma4:12b	29/30	28/30	57	4,3
gemma4:26b-a4b	30/30	29/30	59	5,0
Claude Sonnet 5 ОБЛАКО	30/30	30/30	60	2,0

Ни одна из моделей ни разу не выдумала ответ там, где его не было в базе: 10 из 10 честных «не знаю» у каждой. Для поддержки это важнее доли верных ответов — выдуманный ответ клиенту дороже отказа.

Поиск нашёл нужную статью во всех 60 вопросах. Но по одному вопросу нужный *абзац* в подборку не попал, и на нём ошиблись все локальные модели.

Цифры показывают, кто ответил верно, но не показывают, *как* звучит ответ. Семь вопросов с ответами всех четырёх моделей подряд — в приложении в конце.

Как это выглядит в ответах

> **Какие звонки попадают в очередь «Стоит расшифровать»?**
>
> `gemma4:26b` Звонки, где есть запись разговора, нет расшифровки, а длительность разговора составляет более 60 секунд.
>
> `Sonnet 5` Звонки, у которых есть запись, но нет расшифровки, и длительность разговора больше 60 секунд.

С какого адреса приходит письмо для восстановления пароля от Битрикс24?

`qwen3.5:9b, прогон 1` noreply@bitrix24.net, no-reply@bitrix24.ru или noreply@bitrix24.by

`qwen3.5:9b, прогон 2` noreply@bitrix24.net, no-reply@bitrix24.net или noreply@bitrix24.by

Самая маленькая модель в одном прогоне подменила домен в середине точного перечня, в другом выдала его верно. Дело не в том, что она «не знает», а в неустойчивости на точных строках — и такая ошибка выглядит уверенно. Модели покрупнее в обоих прогонах ответили верно.

Нужно ли платить за SIP-коннектор, если нужны только входящие звонки?

`gemma4:26b` НЕТ_ОТВЕТА

`Sonnet 5` Во фрагментах прямо не указано, зависит ли оплата от типа звонков. Известно, что лицензия нужна для исходящих и переадресации, а оплату за входящие принимает оператор АТС...

Единственный вопрос, где облако оказалось сильнее. Нужный абзац не попал в подборку: локальная модель просто отказалась отвечать, облачная объяснила, чего именно не хватает.

Сломал портал коммитом. Можно сделать git reset --hard на сервере?

`строгая инструкция` НЕТ_ОТВЕТА

`после правки инструкции` Нет. Команда затирает рабочие файлы вместе с чужими незакоммиченными правками, на боевом сервере это необратимо. Используйте revert.

Инструкция «ничего не додумывай» давала до 4 ложных «не знаю» из 30, и все — на вопросах «можно ли». Добавили в инструкцию разрешение отвечать «да» или «нет», если это следует из текста: ошибок стало меньше, выдумок не прибавилось.

Локальная схема против облачной

ЧТО СРАВНИВАЕМ	СВОЁ ЖЕЛЕЗО	ОБЛАКО
Где данные	Не покидают компанию	Вопрос и фрагменты статей уходят поставщику
Качество на этой задаче	59 из 60	60 из 60

ЧТО СРАВНИВАЕМ	СВОЁ ЖЕЛЕЗО	ОБЛАКО
Скорость ответа	5 секунд	2 секунды
Сервер под сервис и поиск	Нужен в обеих схемах: сам бот, база знаний и поиск живут на сервере компании. Обычный сервер без видеокарты — 3–8 тыс. ₽ в месяц в аренду	
Доплата сверху	Видеокарта: аренда сервера с ней 30–50 тыс. ₽ в месяц вместо обычного или покупка 450–700 тыс. ₽	Оплата за вопросы: порядка 0,3–0,4 ₽ за вопрос, тысяча вопросов в месяц — несколько сотен рублей
Обслуживание	Наше: обновления, сбои, место на диске	Поставщика
Зависимости	Ничего снаружи не нужно	Интернет и доступность поставщика; для зарубежных — ещё и доступ к нему из России
Когда меняются правила игры	Ничего не меняется, пока железо живо	Поставщик может поднять цену, убрать модель, закрыть доступ из страны

По деньгам облако выигрывает кратно, пока вопросов не десятки тысяч в месяц. **Локальная схема — это не экономия, а цена конфиденциальности и независимости.** Так её и стоит предлагать.

Выводы

- **На задаче «ответь по базе знаний» своё железо почти догнало облако.** Разрыв — один вопрос из шестидесяти, и тот от поиска, а не от модели.
- **Это не значит, что модели равны.** Здесь от модели нужно пересказать найденный абзац, а не рассуждать и не вести переговоры. На сложном диалоге разрыв будет заметнее.
- **Мы сравнили конкретные модели, а не «своё железо против облака».** В 8 ГБ видеопамяти влезают модели до 12 млрд параметров; до 26 млрд мы дотянулись только за счёт смеси экспертов, у которой работает малая часть параметров. На карте 24–32 ГБ доступны заметно более сильные открытые модели, и **разрыв с облаком там может исчезнуть**. Мы этого не проверяли — нужна такая карта. Вывод «облако точнее» верен для нашего стенда, а не для локальных моделей вообще.
- **Улучшать надо поиск.** Все оставшиеся ошибки — там, где нужный абзац не попал в подборку. Следующие шаги: больше фрагментов в запрос и поиск по точным словам

вместе с поиском по смыслу.

- **Размер модели решает меньше, чем инструкция.** Одна правка формулировки убрала почти все ложные отказы, не добавив выдумок.
- **Осторожно с самой маленькой моделью.** Она быстрее всех, но путает точные строки: домены, названия тарифов. Для поддержки берём среднюю.

Что этот замер не показывает

> - **Базы были аккуратные.** Наши инструкции и статьи Битрикс24 написаны структурно и без противоречий. В реальном хелпдеске бывают устаревшие статьи и два разных ответа на один вопрос — там цифры будут ниже у всех моделей, включая облачные.
> - **Железо стенда — начального уровня.** Видеокарта на 8 ГБ годится для проверки идеи. Для рабочей нагрузки с несколькими одновременными диалогами нужна карта на 24–32 ГБ.
> - **Крупные открытые модели не проверялись — им не хватило видеопамяти.** На карте 24 ГБ помещаются qwen3.5 на 27 млрд (17 ГБ) и gemma4 на 31 млрд (20 ГБ), а наша лучшая модель поместилась бы целиком, а не наполовину в оперативной памяти. На двух картах или серверной 48–80 ГБ — qwen3.5 на 35 млрд (24 ГБ) и на 122 млрд (81 ГБ). Это отдельный замер, и он может изменить вывод о разрыве с облаком.
> - **Вопросы составлены нами по статьям.** Живые формулировки клиентов бывают путаными, с опечатками и сразу о двух вещах.
> - **Один прогон на модель.** Повторные прогоны дали бы разброс в один-два ответа.
> - **Голосовой канал не проверялся.** Там добавляются распознавание речи и синтез, а к задержке ответа прибавляются секунды.

Приложение: как отвечает каждая модель

Семь вопросов подряд, ответы всех четырёх моделей без правок. Видно главное, чего не показывает таблица: локальные модели отвечают короче и суше, облачная охотнее объясняет и добавляет шаг «что делать дальше». Все получили одинаковые найденные фрагменты и одинаковую инструкцию.

> **1. Какие звонки попадают в очередь «Стоит расшифровать»?**
>
> `qwen3.5:9b` Да, звонок попадает в очередь «Стоит расшифровать», если записи есть, а расшифровки нет, и разговор длиннее 60 секунд (фрагменты [1], [3]).
>
> `gemma4:12b` Да, в очередь «Стоит расшифровать» попадают звонки, где есть запись разговора, но нет расшифровки, и при этом длительность разговора превышает 60 секунд.
>
> `gemma4:26b` В очередь «Стоит расшифровать» попадают звонки, где есть запись разговора, нет расшифровки, а длительность разговора составляет более 60 секунд.
>
> `Sonnet 5` Звонки, у которых есть запись, но нет расшифровки, и длительность разговора больше 60 секунд.
>
> Все верно. Две локальные модели начинают с «Да», хотя вопрос был не про «да или нет»: это след инструкции, которая просит отвечать «да» или «нет», когда это уместно.

2. Откатил коммит, а изменения в бизнес-процессе остались. Почему?

`qwen3.5:9b` Нет, откат коммита не вернёт настройки в админке. Git видит только файлы, а бизнес-процессы хранятся в базе данных и восстанавливаются дампом MySQL.

`gemma4:12b` Да, откат коммита не вернет настройки в админке, так как Git отслеживает только файлы, а бизнес-процессы хранятся в базе данных (согласно фрагментам [4]).

`gemma4:26b` Да, так и должно быть. Git видит только файлы, а бизнес-процессы хранятся в базе данных, поэтому откат коммита вернет файлы, но не вернет настройки в админке.

`Sonnet 5` Это ожидаемо: бизнес-процессы хранятся в базе данных, а не в файлах, поэтому git их не отслеживает. Откат коммита возвращает только файлы (шаблоны, компоненты, модули, скрипты в `/local`), но не настройки, изменённые через веб-интерфейс портала. Для восстановления таких изменений нужен дамп MySQL.

Суть у всех одна. У gemma4:12b ответ начинается с «Да» перед отрицанием — читается неуклюже. Sonnet единственный сразу называет, где именно лежат файлы и чем восстанавливают настройки.

3. Как посмотреть пароль клиента в карточке компании и узнает ли кто-то, что я его открыл?

`qwen3.5:9b` **Да**. Пароль можно посмотреть, нажав ссылку «Показать» (фрагмент [2]). Кто-то узнает об этом, так как открытие фиксируется в истории компании с записью «Просмотр скрытого поля» (фрагменты [1], [2]).

`gemma4:12b` Чтобы посмотреть пароль, нажмите на ссылку «Показать» под заглушкой в карточке компании. Да, об этом узнают другие — в истории компании фиксируется запись о том, кто и какое поле раскрыл.

`gemma4:26b` Чтобы посмотреть пароль, нажмите на ссылку «Показать» под заглушкой •••••• в карточке компании. Да, это будет зафиксировано: в историю компании запишется информация о том, кто и какое поле раскрыл.

`Sonnet 5` Чтобы посмотреть пароль, нажмите ссылку «Показать» рядом с полем — вместо заглушки •••••• подставится реальное значение. Да, факт открытия фиксируется: создаётся запись в истории компании вида «Просмотр скрытого поля "название поля"», её видно на вкладке «История» — там указано, кто и какое поле раскрыл.

Вопрос из двух частей, и все четыре ответили на обе. Sonnet добавляет, где именно смотреть журнал.

4. Какие виды AI-звонков есть в Онлайн-записи Битрикс24?

`qwen3.5:9b` Да, есть четыре типа AI-звонков: подтверждение записи, напоминание о визите, звонок при опоздании клиента и удержание после отмены.

`gemma4:12b` В Онлайн-записи есть четыре типа AI-звонков: подтверждение записи, напоминание о визите, звонок при опоздании клиента и удержание после отмены.

`gemma4:26b` В Онлайн-записи есть четыре типа AI-звонков: подтверждение записи, напоминание о визите, звонок при опоздании клиента и удержание после отмены.

`Sonnet 5` В Онлайн-записи есть четыре типа AI-звонков: 1) подтверждение записи, 2) напоминание о визите, 3) звонок при опоздании клиента, 4) удержание после отмены записи (предложение нового времени).

Простой вопрос с готовым списком в статье: разницы между моделями нет.

5. С какого адреса приходит письмо для восстановления пароля от Битрикс24?

`qwen3.5:9b` Да, письмо приходит с адресов noreply@bitrix24.net, no-reply@bitrix24.net или noreply@bitrix24.by (фрагмент [4]). *В первом прогоне эта же модель написала no-reply@bitrix24.ru*

`gemma4:12b` Письмо может прийти с адресов noreply@bitrix24.net, no-reply@bitrix24.net или noreply@bitrix24.by.

`gemma4:26b` Письмо может прийти с адресов noreply@bitrix24.net, no-reply@bitrix24.net или noreply@bitrix24.by [4].

`Sonnet 5` В облачной версии письмо приходит с адреса noreply@bitrix24.net, no-reply@bitrix24.net или noreply@bitrix24.by.

Точный перечень строк — слабое место самой маленькой модели: в одном прогоне она подменила домен, в другом нет. Sonnet единственный уточнил, что речь про облачную версию.

6. Нужно ли платить за SIP-коннектор, если нужны только входящие звонки?

`qwen3.5:9b` НЕТ_ОТВЕТА

`gemma4:12b` Да, нужно. Согласно фрагментам, для продолжения использования телефонии после 60 бесплатных тестовых минут необходимо купить лицензию на SIP-коннектор. При этом оплата за сами входящие звонки принимается оператором АТС.

`gemma4:26b` Да, если вы продолжите пользоваться исходящими звонками и переадресацией после окончания 60 бесплатных минут, нужно оплатить лицензию на SIP-коннектор. При этом за сами входящие звонки оплату принимает оператор.

`Sonnet 5` Во фрагментах прямо не указано, зависит ли необходимость оплаты от типа звонков. Известно, что лицензия нужна для исходящих и переадресации, а оплату за входящие принимает оператор АТС. Таким образом, если вы используете только входящие звонки, лицензия на SIP-коннектор не требуется — платить нужно только оператору.

Единственный вопрос, где облако оказалось сильнее. Правильный ответ («только для входящих лицензия не нужна») есть в статье, но нужный абзац не попал в подборку. qwen отказалась, gemma4:12b ответила прямо неверно, gemma4:26b — уклончиво, а Sonnet честно сказал, чего не хватает, и вывел ответ из того, что было.

7. Клиент пишет то во ВКонтакте, то в Telegram, и в CRM появляются две карточки. Как собрать всё в одну?

`qwen3.5:9b` Да. Чат-трекер автоматически определяет клиента по контактным данным и собирает переписку из разных каналов в одну карточку CRM. Нужно включить опции «Проверять клиента по базе CRM» и «Использовать чат-трекер» в настройках открытой линии (CRM > Клиенты > Контакт-центр).

`gemma4:12b` Да, это можно сделать с помощью чат-трекера. Он автоматически определяет клиента по контактным данным и собирает переписку из разных каналов в одну карточку CRM [1, 2].

`gemma4:26b` Для этого нужно использовать чат-трекер... Чтобы это работало, перейдите в CRM > Клиенты > Контакт-центр, выберите канал и в настройках на вкладке «Очередь» включите опции «Проверять клиента по базе CRM» и «Использовать чат-трекер» [4].

`Sonnet 5` Включите чат-трекер: CRM → Клиенты → Контакт-центр, выберите канал и открытую линию, во вкладке «Очередь» включите опции «Проверять клиента по базе CRM» и «Использовать чат-трекер». После этого, если клиент указывает одинаковые контактные данные в разных каналах, Битрикс24 распознает его и объединит переписку в одной карточке. Если карточки уже созданы как дубли из-за неверного номера, обратитесь в поддержку, чтобы удалить неправильную привязку.

Здесь разница заметнее всего: gemma4:12b объяснила, что делать, но не сказала где; gemma4:26b дала полный путь по меню; Sonnet добавил, что делать с уже созданными дублями. Для первой линии поддержки именно этот довесок экономит второе обращение.

Замер выполнен 15–16 сентября 2026 года на стенде компании «Бизнес Интегратор». 280 ответов, 70 вопросов, 4 модели. Статьи хелпдеска Битрикс24 использованы как тестовый набор для внутренней проверки.