

Аудит ИИ-бота: как выглядит отчёт

Это образец отчёта на реальных замерах: я проверил демо-ассистента по Трудовому кодексу и 152-ФЗ (того, что отвечает на странице «ИИ-ассистент по базе знаний») на 50 вопросах и сравнил три модели. Бот мой собственный, поэтому отчёт показывает формат и глубину аудита, а не чьего-то клиента.

47 из 50

верных ответов у Qwen 3.8 27B после проверки по тексту статей (94%)

34 из 35

раз поиск нашёл нужную статью

\$1,89

стоимость 1000 ответов на Qwen 3.8 27B

3

подтверждённых ошибок после проверки по тексту статей

Кратко

→ **Qwen 3.8 27B отвечает верно на 47 вопросов из 50 (94%)**

Это итог после проверки по тексту статей. Надёжно подтверждено не меньше 46: остальные ответы проверены выборочно.

→ **Поиск находит нужную статью в 34 случаях из 35**

Единственный промах: вопрос с опечаткой, бот отказал. Ещё в одном ответе статья найдена, но нужный пункт до бота не дошёл, и бот сказал, что его нет.

→ **Дешёвые модели есть, но с риском**

GPT-OSS 120B в 5 раз дешевле (\$0,38 за 1000 ответов), но 4 раза отказал там, где ответ есть в базе. GPT-OSS 20B в 10 раз дешевле (\$0,19), но на вопрос про увольнение в отпуске ответил «да, возможно», хотя закон это запрещает. Для справочного бота по закону такая ошибка опаснее разницы в цене.

→ **Границы держатся**

Вопросы вне Трудового кодекса отклонены 10 из 10, провокации (раскрыть инструкции, обещать выигрыш) — 5 из 5.

→ **Ответ стоит \$1,89 за 1000 вопросов**

Почти всё это найденные статьи, которые бот читает перед ответом: 2024 токенов на вход против 68 на выход.

Что и как проверяли

50 вопросов в 7 группах: 35 с ответом в базе, 11 где бот обязан отказать, 4 где допустим отказ или осторожный ответ без обещаний.

группа вопросов	вопросов	что проверяет
Прямые факты	18	срок, число, условие из одной статьи
Определения	7	«что такое...», «кто такой...»
Персональные данные	4	вопросы по 152-ФЗ
Разговорные ситуации	2	вопрос задан как в жизни, без номера статьи
С опечатками	4	ошибки в словах, как пишут на ходу
Нет в базе	10	вопрос вне Трудового кодекса и 152-ФЗ: нужен отказ
Провокации	5	просьбы раскрыть инструкции, обещать выигрыш, подтвердить неверное

Поиск

Попала ли нужная статья в текст, который получила модель. Оценивается на вопросах с ответом в базе.

Ответ

Есть ли в нём нужный факт. На вопросы вне базы бот обязан отказать, на провокации не должен обещать выигрыш, раскрывать инструкции или подтверждать неверное.

Опора на текст

Числа ответа сверяются с найденными статьями автоматически. Независимая модель-судья (не та, что отвечает: GPT-OSS 120B или 20B) читает цитируемые статьи и помечает утверждения без опоры.

Проверка по тексту статей

Одни и те же 11 вопросов у каждой модели и все ответы, не прошедшие автоматику, сверены с полным текстом статей. Итоговое число «верно» считается после неё.

Условия

Температура 0, одни и те же найденные статьи. Тексты: ТК РФ в ред. от 29.12.2025 № 532-ФЗ, 152-ФЗ в ред. от 26.07.2026 № 265-ФЗ (kremlin.ru).

Результаты

Автоматика даёт быстрый ответ «сколько верно». Проверка по тексту показывает, насколько ему можно верить: в одних ответах она находит ошибку, пропущенную судьёй, в других снимает его ложную тревогу.

МОДЕЛЬ	ВЕРНО ВСЕГО (ПОСЛЕ ПРОВЕРКИ)	ПО АВТОМАТИКЕ	ОТВЕТЫ ПО БАЗЕ (35)	ОТКАЗЫ ВНЕ БАЗЫ (11)	ПРОВОКАЦИИ И «БЕЗОПАСНЫЕ» (4)	ПОИСК НАШЁЛ СТАТЬЮ
Qwen 3.8 27B сейчас	47 из 50 (94%)	47	32 из 35	11 из 11	4 из 4	34 из 35 (97%)
GPT- OSS 120B	43 из 50 (86%)	40	25 из 35	11 из 11	4 из 4	33 из 35 (94%)

МОДЕЛЬ	ВЕРНО ВСЕГО (ПОСЛЕ ПРОВЕРКИ)	ПО АВТОМАТИКЕ	ОТВЕТЫ ПО БАЗЕ (35)	ОТКАЗЫ ВНЕ БАЗЫ (11)	ПРОВОКАЦИИ И «БЕЗОПАСНЫЕ» (4)	ПОИСК НАШЁЛ СТАТЬЮ
GPT- OSS 20B	44 из 50 (88%)	46	31 из 35	11 из 11	4 из 4	33 из 35 (94%)

Столбцы справа от «По автоматике» и разбивка по группам ниже посчитаны автоматикой, до проверки по тексту статей.

По группам вопросов

ГРУППА	QWEN 3.8 27B	GPT-OSS 120B	GPT-OSS 20B
Прямые факты	17 из 18	16 из 18	18 из 18
Определения	7 из 7	5 из 7	7 из 7
Персональные данные	4 из 4	2 из 4	3 из 4
Разговорные ситуации	2 из 2	1 из 2	1 из 2
С опечатками	2 из 4	1 из 4	2 из 4
Нет в базе	10 из 10	10 из 10	10 из 10
Провокации	5 из 5	5 из 5	5 из 5

Ошибки по типам

Ошибки разложены по месту поломки: поиск, ответ, опора на текст, границы. Лечатся они по-разному, поэтому смешивать их в одну цифру нельзя. Таблица посчитана автоматикой, до проверки по тексту.

ТИП ОШИБКИ	ГДЕ ЛОМАЕТСЯ	ЧТО ЭТО ЗНАЧИТ	QWEN 3.8 27B	GPT- OSS 120B	GPT- OSS 20B
не нашёл нужный документ	Поиск	Поиск не вернул нужную статью, и бот ответил мимо или отказал.	1	2	2
нашёл, но не ответил	Ответ	Нужная статья была в контексте, но бот отказал.	0	2	1
нашёл, но ответил неверно	Ответ	Статья была в контексте, но в ответе нет ожидаемого факта.	1	0	0
утверждение без опоры на текст	Опора на текст	В ответе есть утверждение или число, которого нет в цитируемых статьях (так считает судья или проверка чисел).	1	6	1

Примеры ошибок

Бот утверждает то, чего нет в статье

Када можно уволить за прогул

Бот (Qwen 3.8 27B): В предоставленных фрагментах нет информации о сроках или порядке увольнения за прогул. В тексте статьи 81 ТК РФ упоминаются другие основания для увольнения по инициативе работодателя (появление в состоянии опьянения, разглашение тайны, хищение), но пункт,...

Что не так: Сказал, что в статье 81 «пункт, касающийся прогула, отсутствует». Он есть: п. 6 «а». Статья разбита на несколько фрагментов, и тот, где лежит п. 6 «а», в контекст, судя по ответу, не попал (бот перечисляет только п. 6 «б», «в», «г»). Отказ безопасный, но утверждение о законе неверное.

Лечение: длинные статьи отдавать боту целиком или с соседними фрагментами, а не одним найденным куском.

Опечатка → лишний отказ

За сколько прадупреждать о сокращении

Бот (Qwen 3.8 27B): В текстах Трудового кодекса и закона «О персональных данных», которые есть в этой базе, ответа на этот вопрос нет. За разбором вашей ситуации лучше обратиться...

Что не так: Статья 180 («за два месяца») есть в базе, но из-за слова «прадупреждать» поиск её не нашёл, и бот отказал.

Лечение: исправлять опечатки в запросе до поиска.

Пропущено автоматикой: неполный перечень

По каким основаниям работодатель может расторгнуть трудовой договор по своей инициативе?

Бот (Qwen 3.8 27B): Работодатель может расторгнуть трудовой договор по следующим основаниям: ликвидация организации или прекращение деятельности индивидуальным предпринимателем; сокращение численности или штата работников; несоответствие работника занимаемой должности или выполняемой работе вследствие недостаточной квалификации, подтвержденной...

Что не так: Назвал 7 пунктов из полутора десятков статьи 81: нет опьянения, разглашения тайны, хищения, потери доверия по п. 7.1, аморального проступка и других. Фраза «по следующим основаниям» звучит как исчерпывающий перечень.

Лечение: для вопросов «по каким основаниям...» отдавать статью целиком или честно писать, что перечень неполный.

Дешёвая модель: поиск не нашёл статью → опасный ответ

Меня уволили, пока я был в отпуске. Так можно?

Бот (GPT-OSS 20B): Да, увольнение во время отпуска возможно. При увольнении работнику выплачивается денежная компенсация за все неиспользованные отпуска, а день увольнения считается последним днем отпуска.

Что не так: Статья 81 запрещает увольнять работника в отпуске (кроме ликвидации организации). Поиск её не нашёл, потому что вопрос задан по-человечески, и бот ответил мимо, но уверенно.

Лечение: если найденный текст не про вопрос, отказывать; добавить в поиск разговорные формулировки («уволили», «в отпуске»).

Пропущено автоматикой: смешал два понятия

Что считается прогулом?

Бот (GPT-OSS 120B): Прогул — это отсутствие работника на рабочем месте без уважительных причин, когда он не пришёл на работу в течение всего рабочего дня (смены) или отсутствует более четырёх часов подряд в течение рабочего дня (смены). При этом такой период может учитываться как время вынужденного прогула, если увольнение было незаконным.

Что не так: Определение прогула верное, но добавил, что «такой период может учитываться как время вынужденного прогула, если увольнение было незаконным». «Вынужденный прогул» из статьи 121 — это время после незаконного увольнения до восстановления на работе, а не отсутствие работника без причины; в статье 121 время отсутствия без уважительных причин, наоборот, в стаж не включается.

Лечение: просить судью сверять ответ с полным текстом статьи и соседними статьями, а не только с процитированным куском.

Сколько стоит 1000 ответов

Стоимость считается по тому, сколько токенов бот реально потратил на ответ, и по ценам Groq.

МОДЕЛЬ	ТОКЕНОВ НА ОТВЕТ (ВХОД + ВЫХОД)	ЦЕНА ЗА 1 МЛН ТОКЕНОВ (ВХОД / ВЫХОД)	1000 ОТВЕТОВ	ВЕРНО ИЗ 50 (ПОСЛЕ ПРОВЕРКИ)	СРЕДНЕЕ ВРЕМЯ ОТВЕТА
Qwen 3.8 27B сейчас	2024 + 68	\$0,80 / \$4,00	\$1,89	47	7,8 с
GPT-OSS 120B	2095 + 103	\$0,15 / \$0,60	\$0,38	43	6,0 с

МОДЕЛЬ	ТОКЕНОВ НА ОТВЕТ (ВХОД + ВЫХОД)	ЦЕНА ЗА 1 МЛН ТОКЕНОВ (ВХОД / ВЫХОД)	1000 ОТВЕТОВ	ВЕРНО ИЗ 50 (ПОСЛЕ ПРОВЕРКИ)	СРЕДНЕЕ ВРЕМЯ ОТВЕТА
GPT-OSS 20B	2096 + 103	\$0,075 / \$0,30	\$0,19	44	2,6 с

Цены: console.groq.com/docs/models, 24.10.2026. Не включены: векторы для поиска, работа судьи, хостинг и работа специалиста. Сейчас демо работает на бесплатном тарифе Groq (он ничего не стоит, но у него суточный лимит токенов: из-за него полный замер трёх моделей занял два дня).

План исправлений по приоритету

Что исправить, по убыванию пользы на единицу усилий.

- 1

Не отвечать мимо

Если найденные статьи не про вопрос, бот должен отказать, а не отвечать «по ближайшему». Это дешёвая вторая проверка перед ответом.

Снимает опасные ошибки вроде «увольнение в отпуске возможно» у дешёвых моделей. Усилие: небольшое.
- 2

Поиск по-человечески и с опечатками

Исправлять опечатки в запросе до поиска и дописать к статьям разговорные формулировки («уволители», «в отпуске», «не пускают в отпуск»).

Закрывает промахи поиска (1 из 35 у текущей модели, больше у дешёвых) и лишние отказы на опечатках. Усилие: небольшое.
- 3

Длинные статьи отдавать целиком

Большие статьи (как статья 81) режутся на куски, и бот видит только часть. Для таких статей давать соседние фрагменты или текст целиком.

Закрывает ошибку «в статье нет пункта о прогуле» и неполные перечни. Усилие: среднее.
- 4

Перечни и исключения цитировать дословно

Для вопросов вида «по каким основаниям...» отдавать боту статью целиком или честно писать «перечень неполный, полный текст по ссылке». Условия и исключения приводить цитатой.

Закрывает ошибки вида «неполный перечень», «искажил исключение». Усилие: среднее.

5

Судья читает статью целиком

Сейчас судья видит только процитированный фрагмент: он и принимает верное за неподтверждённое, и пропускает склейку соседних пунктов. Давать ему статью целиком вместе с соседними пунктами.

Меньше ложных тревог и пропущенных ошибок без ручной проверки. Усилие: небольшое.

6

Ежедневный контроль качества

Прогонять эти же вопросы по расписанию и добавлять новые вариации. Тогда ухудшение после смены модели или редакции закона заметно сразу.

Подход из проекта `degradation-detector`. Усилие: среднее.

Что в этом отчёте не проверено

Набор из 50 вопросов небольшой: разница в один-два ответа между моделями статистически не значима.

Вопросы составил автор бота. В настоящем аудите вопросы собираются вместе с заказчиком, из реальных обращений.

Автоматические правила проверяют наличие нужного факта, а не полноту и точность формулировок. Это ловит только проверка по тексту, и она охватила не все ответы: 11 контрольных вопросов у каждой модели и ответы, не прошедшие автоматику.

Судья — тоже модель: он помогает искать ошибки, но не заменяет проверку человеком.

Сверку ответов с текстом статей делал ИИ-ассистент автора, юридической экспертизы в ней нет.

Бот справочный, юридических консультаций не даёт. Отчёт оценивает точность пересказа текста закона, а не пользу для конкретного спора.