Вскрытие AI-решения. Уровень 3.

Независимый технический аудит бота «AI-Pro Solutions»
в сети ресторанов «Хлеб & Соль»

Это пример. Под ваш случай я делаю такое же вскрытие за 50–80 000 ₽ — в зависимости от сложности.

Максим Чирков · AI-аудитор
chirkov.tech · @on_max · май 2026

Что у клиента есть и в чём сомнения

Управляющий сети «Хлеб & Соль» заказал независимое вскрытие бота, купленного 3 месяца назад.

Что куплено

Подрядчик	ООО «AI-Pro Solutions» (название изменено)
Дата запуска	3 месяца назад
Цена внедрения	280 000 ₽ единоразово
Поддержка	35 000 ₽/мес
За 3 месяца уплачено	385 000 ₽

Заявленный функционал

- Автоматические ответы на отзывы (2ГИС, Я.Карты, Google)
- AI-ассистент в WhatsApp для типовых обращений
- «Уникальная нейросеть, обученная на данных вашего ресторана»
- Интеграция с CRM
- Аналитика по типам обращений

Что не нравится клиенту

× 2 публичных скандала: ответы на негативные отзывы получились «издевательскими»
× Гости в WhatsApp жалуются, что «робот ничего не понимает»
× На вопросы про меню бот выдумывает несуществующие блюда (галлюцинации)
× Маркетолог не понимает, что значит «обученная на наших данных» — учебных данных не присылали
× Подрядчик месяцами «улучшает» — ощутимых изменений нет

Запрос на аудит

Что реально внутри? Соответствует обещанному? Адекватная ли цена? Чинить или выбрасывать? Если переходить на новое — что взять и за сколько?

Резюме вскрытия

Главное в одной странице. Подробности — на следующих 15 страницах.

160 000 ₽ + 300 000 ₽/год

Переплата клиента подрядчику AI-Pro Solutions

160к — единоразово на внедрении · 25к/мес × 12 = 300к/год на поддержке

Что реально внутри	GPT-3.5-turbo (модель 2022 г., устаревшая) через простую обёртку. «Уникальная нейросеть» — ложь.
Сложность решения	Низкая. Прототип такого функционала собирается за 2–3 рабочих дня.
Справедливая цена внедрения	80 000 – 120 000 ₽ вместо 280 000
Справедливая поддержка	8 000 – 12 000 ₽/мес вместо 35 000
Критические проблемы	Галлюцинации (выдумывает блюда), нет вычитки негатива, утечка ПДн в OpenAI, открытый S3-бакет

Вердикт

ВЫБРАСЫВАТЬ

Чинить технически можно, но дешевле собрать заново на правильной архитектуре.

Методология вскрытия

Как я работаю с уже внедрённым решением — 8 рабочих дней.

День 1 — доступы и документы
Запрашиваем у подрядчика техническую документацию, доступы к админке, договор, переписку.

Дни 2–3 — чёрный ящик
Прогоняем через бот ~100 типовых и краевых запросов, фиксируем ответы. По ответам определяем модель.

День 4 — архитектура
Если есть доступ — смотрим код/конфиги. Если нет — реверс-инжинирим через поведение.

День 5 — промпты
Пытаемся достать системный промпт (известные техники jailbreak — законно для аудита своей системы).

День 6 — безопасность
Куда уходят данные, есть ли утечки ПДн, в каком виде хранится история диалогов.

День 7 — договор и SLA
Что подрядчик гарантировал и что реально делает. База для юридической позиции.

День 8 — сводный отчёт
Что внутри, сколько стоит на рынке, чинить/выбрасывать — с цифрами.

Итого: 8 рабочих дней. Цена работ: 50 000 – 80 000 ₽ в зависимости от объёма реверс-инжиниринга.

Заявления подрядчика vs реальность

Что они продали и что внутри.

Заявление подрядчика	Реальность
«Уникальная нейросеть»	Стандартный GPT-3.5-turbo через OpenAI API
«Обучена на ваших данных»	Few-shot: 8 примеров впихнуты в системный промпт. Это НЕ обучение.
«Понимает контекст вашего меню»	Меню текстом в промпте. Любое обновление меню — ручной апдейт подрядчиком.
«Глубокая интеграция с CRM»	Webhook на изменение статуса заявки. 30 минут работы программиста.
«AI-аналитика обращений»	Категоризация по ключевым словам через тот же GPT. Простой скрипт.
«24/7 мониторинг качества»	Логи пишутся, но никто их не смотрит. Скандалы возникли через 2 недели после факта.

⚠ Критический пункт. «Защищённое хранение данных» — логи диалогов лежат в открытом S3-бакете. Доступны без авторизации. Это утечка персональных данных гостей. Подробности — на стр. 11.

Технический разбор

Какая модель внутри и как устроена архитектура.

Как определили модель

Прогнали 30 тестов, на которые GPT-3.5, GPT-4, Claude и YandexGPT дают характерно разные ответы. Скорость ответа (1.2 сек), стилистика ошибок, лимит контекста (~4000 токенов в ответах) — однозначно GPT-3.5-turbo.

Архитектура

Бот	Простая обёртка вокруг OpenAI API
Системный промпт	Один статичный шаблон, ~800 токенов
База знаний (меню, политики)	Впихнута в промпт текстом, без RAG
Память диалога	Последние 4 сообщения
Интеграция с CRM	Один webhook
Аналитика	Ежедневный скрипт, GPT раскидывает обращения по категориям

Сложность всей системы: уровень джуниора-разработчика, 2–3 дня работы.

Чего нет, хотя должно быть

- ✕ RAG для меню (чтобы при обновлении меню бот сразу знал новое)
- ✕ Эскалация на человека для негативных отзывов
- ✕ Вычитка ответов перед публикацией
- ✕ Гарантия защиты данных (корпоративный API или локальная модель)
- ✕ Отдельный промпт для разных каналов (2ГИС vs WhatsApp требуют разного тона)
- ✕ Логирование с алертами на странные ответы

Разбор промптов

Системный промпт, который удалось достать через jailbreak-технику.

```
Ты — AI-ассистент сети ресторанов "Хлеб & Соль".
Отвечай вежливо и дружелюбно.
Если жалуются — извинись. Если хвалят — поблагодари.

[Далее идёт текст меню всех 3 ресторанов на 4 страницы]

[8 примеров ответов в стиле "вопрос-ответ"]

Не используй грубые слова. Не давай скидок.
Не обещай ничего кроме того, что написано в меню.
```

Что с этим не так

1. **Промпт примитивный.** Никакой роли (тон, ценности компании), нет правил эскалации, нет инструкции «если не знаешь — скажи 'не знаю'».

2. **Меню в промпте текстом.** Модель не понимает структуру, плавает в названиях, иногда смешивает блюда из разных ресторанов.

3. **8 few-shot примеров — мало.** Плюс примеры однотипные (только позитивные кейсы).

4. **Нет защиты от галлюцинаций.** Отсюда выдумки блюд («Маленький кролик» — придумано ботом).

5. **Нет инструкций по негативу.** Отсюда «издевательские» ответы на 1-звёздочные отзывы.

Качество ответов

Прогнали 100 типовых запросов. На простом — работает; на сложном и негативе — проваливается.

Категория	Тестов	Результат	Комментарий
Время работы / адрес	20	🟢 18 🟡 2	Простое работает
Меню — стандартные вопросы	20	🟢 12 🟡 5 🔴 3	3 галлюцинации
Меню — нестандартные («без глютена?»)	10	🟢 4 🟡 4 🔴 2	2 галлюцинации
Положительные отзывы — ответ	15	🟢 10 🟡 5	OK
Нейтральные отзывы — ответ	15	🟢 8 🟡 7	OK
Негативные отзывы — ответ	10	🟢 2 🟡 3 🔴 5	**5 неприемлемых**
Бронирование столика	10	🟢 3 🟡 5 🔴 2	Не передал в CRM × 2

Конкретные провалы (из логов)

1. Гость пожаловался на холодный суп → бот ответил: «Возможно, вы заказали слишком давно. Спасибо за обратную связь!»

2. Вопрос про веганские опции → бот предложил «Котлету по-киевски без курицы».

3. Гость спросил про детское меню → бот придумал блюдо «Маленький кролик», которого нет.

Что подрядчик продал лишнего

Построчный разбор коммерческого предложения и реальной стоимости работ.

Позиция	Цена в КП	Реальная цена рынка	Что фактически сделано
Разработка уникальной нейросети	120 000 ₽	0 ₽	Подмена понятий (готовый GPT-3.5)
Обучение модели на ваших данных	60 000 ₽	0 ₽	Подмена понятий (нет обучения)
Интеграция с CRM	40 000 ₽	8 000 ₽	1 webhook за день
Настройка ассистента в WhatsApp	30 000 ₽	15 000 ₽	Подключение через Twilio
Система аналитики обращений	30 000 ₽	10 000 ₽	Скрипт с GPT-категоризацией
ИТОГО заявлено	**280 000 ₽**	**33 000 ₽**	**Переплата ~160 000 ₽**

Адекватная цена при честной упаковке

С учётом тестирования, документации, обучения команды: 80 000 – 120 000 ₽. Переплата на внедрении: 160 000 – 200 000 ₽.

По поддержке 35 000 ₽/мес

Реальный объём работ: мониторинг, мелкие правки промптов, расходы на OpenAI API (~3 000 ₽/мес). Адекватная цена: 8 000 – 12 000 ₽/мес. Переплата ~25 000 ₽/мес = 300 000 ₽/год.

Анализ договора и SLA

От этого зависит, можно ли требовать возврат денег.

Что есть в договоре

- ✓ Предмет: «разработка и сопровождение AI-ассистента»
- ✓ Срок внедрения: 30 рабочих дней
- ✓ Гарантия работоспособности: 6 месяцев

Чего НЕТ (и должно было быть)

- ✕ Описание используемых моделей (можно «обновлять» втихую)
- ✕ Метрики качества (точность, доля галлюцинаций, доля эскалаций)
- ✕ SLA по времени реакции на инциденты
- ✕ Гарантия защиты персональных данных
- ✕ Условия передачи доступов при расторжении
- ✕ Запрет на использование ваших данных для обучения других клиентов

Юридическая позиция

Можно требовать расторжения по основанию «несоответствие фактически оказанной услуги заявленному в КП». Шансы выиграть в суде высокие, но дешевле договариваться. Рекомендуем привлечь юриста.

Критические проблемы безопасности

Срочные риски — действовать в течение 24 часов.

⚠ 1. Утечка ПДн через OpenAI
Все диалоги уходят в OpenAI API без специальных условий → используются для обучения моделей. Имена, телефоны, адреса гостей попадают в логи OpenAI. Решение: переход на корпоративный план (Enterprise/API без обучения) или Anthropic/Yandex.

⚠ 2. Открытый S3-бакет с логами
Логи всех диалогов в облачном хранилище без авторизации. Любой человек, знающий URL, скачивает все переписки. На момент аудита (15.05.2026) бакет содержит ~85 000 диалогов за 3 месяца. Нарушение 152-ФЗ. Штрафы за утечку ПДн: 300 000 – 18 000 000 ₽. Срочно: требовать закрыть бакет в течение 24 часов.

⚠ 3. Нет шифрования при передаче
Часть webhook'ов идёт по HTTP, не HTTPS. Возможен перехват.

⚠ 4. Подрядчик имеет доступ ко всему
Если расторгнуть договор — подрядчик может удалить или продать данные клиента. В договоре это не запрещено.

Чинить или выбрасывать

Три варианта на выбор. Рекомендация — третий.

Вариант А — оставить как есть

Цена: 35 000 ₽/мес. Риск: продолжение скандалов, утечка ПДн, штрафы по 152-ФЗ. Не рекомендуется.

✕ Вариант Б — чинить с тем же подрядчиком

Что нужно: расширить системный промпт, добавить RAG для меню, эскалацию негатива на человека, безопасное хранение, корпоративный API. Объём работ для подрядчика: ~80 000 ₽ доплаты (хотя должны делать бесплатно по гарантии). Скорее всего откажется — это переделка большей части их работы.

Не рекомендуется. Даже если согласятся — доверия к их компетенции после вскрытия нет.

 Вариант В — выбросить и пересобрать заново

Время: 2 недели. Цена пересборки: 80 000 – 120 000 ₽ единоразово. Поддержка: 8 000 – 12 000 ₽/мес.

Что получите:
• Современная модель (Claude 3.5 / GPT-4o) с защитой от галлюцинаций
• RAG для меню (автообновление при изменениях)
• Эскалация всех негативных отзывов на менеджера
• Безопасное хранение в РФ-юрисдикции
• Корпоративный API с гарантией неиспользования данных
• Метрики и алерты

Рекомендуется.

Экономика перехода

Текущие затраты	35 000 ₽/мес = 420 000 ₽/год
Новое решение	12 000 ₽/мес = 144 000 ₽/год + 100 000 ₽ разовых
Экономия в первый год	176 000 ₽
Экономия со второго года	276 000 ₽/год

Если решено пересобрать — план работ

2 недели до запуска. По неделям.

Неделя 1 — фундамент

День 1–2. Дизайн архитектуры, выбор модели (Claude 3.5 Sonnet через корпоративный API)

День 3–4. RAG для меню — векторная база с автообновлением

День 5. Системный промпт с правилами эскалации и защитой от галлюцинаций

День 6–7. Интеграции (WhatsApp, 2ГИС, Я.Карты, CRM) на новых, защищённых каналах

Неделя 2 — качество и запуск

День 8–9. 200 тестов качества, доводка промптов

День 10. Настройка эскалации негативных отзывов на менеджера

День 11. Дашборд метрик (точность, % галлюцинаций, время ответа)

День 12. Обучение команды, передача доступов

День 13–14. Hypercare, мониторинг первых обращений

Чек-лист «как должно быть»

- ✓ Современная модель (Claude 3.5 / GPT-4o / GigaChat Pro)
- ✓ RAG, не статичное меню в промпте
- ✓ Защита от галлюцинаций («не знаю — скажи 'не знаю'»)
- ✓ Эскалация негатива на человека
- ✓ Логи в защищённом хранилище в РФ-юрисдикции
- ✓ Корпоративный API без использования данных для обучения
- ✓ Метрики: точность ≥ 95%, галлюцинаций ≤ 1%, время реакции < 2 сек
- ✓ Документация и SLA с метриками

Карта рисков продолжения работы с подрядчиком

Что может пойти не так, если не двигаться с места.

Риск	Вероятность	Тяжесть	Митигация
Очередной репутационный скандал	● Высокая	● Высокая	Только выход на новое решение
Штраф по 152-ФЗ за S3-бакет	● Средняя	● Высокая	Срочно требовать закрыть, готовить юриста
Подрядчик отказывается передавать доступы	● Высокая	● Высокая	Юрист, договор, параллельная сборка
Подрядчик удаляет/продаёт данные	● Средняя	● Высокая	Запросить копию логов до расторжения
Подорожание GPT API	● Низкая	● Средняя	Новое решение позволяет выбирать модель

Что НЕ автоматизируем даже в новом решении

Принципиальные ограничения. Не должны меняться.

✕ Ответы на негативные отзывы — только через человека
AI готовит черновик с пометкой «не для прямой отправки». Менеджер вычитывает и отправляет.

✕ Финансовые обещания (скидки, компенсации) — только человек
Никогда AI. Даже черновика не делаем.

✕ Решения по жалобам на сотрудников — только человек
AI даже не комментирует.

✕ Обещания, которые могут пойти в суд — только под подпись
AI не даёт никаких гарантий от лица заведения.

Принцип: AI — помощник, не носитель ответственности.

Что вы получаете после Уровня 3

Чек-лист артефактов вскрытия.

✓	**Этот PDF**	Разбор по 8 направлениям
✓	**100 проведённых тестов с ответами**	Excel — что спрашивали, что бот отвечал, как оценили
✓	**Извлечённый системный промпт подрядчика**	Полный текст с разметкой найденных проблем
✓	**Карта проблем с приоритетами**	Что критично, что терпит
✓	**Чек-лист «как должно быть»**	Для нового решения — что обязательно, что желательно
✓	**Юридическое резюме**	Можно идти к юристу
✓	**Запись 2 защитных созвонов с подрядчиком**	Если согласитесь — мы участвуем как ваш техэксперт
✓	**2 недели поддержки в Telegram**	Любые вопросы по переговорам с подрядчиком

Дальнейшие шаги

После Уровня 3 у вас полная картина. Дальше — на выбор.

Вариант А — всё делаете сами — **включено в Уровень 3**

С нашим разбором идёте к подрядчику, требуете возврата/переделки. Параллельно ищете нового исполнителя. Я доступен в TG для консультаций 2 недели.

Вариант Б — пересобираем мы (Уровень 4) — **80 000 – 120 000 ₽**

2 недели работ + поддержка 12 000 ₽/мес. Гарантия экономии: 176 000 ₽ за первый год vs текущее решение.

Вариант В — участвуем как ваш техэксперт на переговорах — **30 000 ₽ за созвон**

Часто только наше присутствие — повод для возврата 50–100% денег. Подрядчики не любят независимых техэкспертов.

О консультанте

Максим Чирков — технический AI-аудитор.

Экс-МТС — выстраивал процессы с нуля. Экс-Т-Банк — оптимизировал работающие. Сейчас помогаю малому и среднему бизнесу разбираться, что им продали под видом AI.

«Мне можно доверять разбор: я работал и на стороне заказчика, и на стороне исполнителя. Знаю, где подрядчики чаще всего наценивают, и как они отмазываются. На моих созвонах подрядчики обычно сразу становятся сговорчивее.»

Контакты

Telegram-канал	«Раздеваю AI» — @razdevayu_ai
Личный контакт	@on_max
Сайт	chirkov.tech
Запись на вскрытие	chirkov.tech или в личку @on_max

Подозреваете подрядчика? Закажите вскрытие. 50–80 000 ₽ · chirkov.tech · @on_max