Как внедрить AI и удержать его на коротком поводке

Опыт внедрения, ошибки
и найденные решения

Мир до LLM — эра Regex

Традиционные роботы строились на жестких скриптах, шаблонах и ключевых словах. Каждый сценарий диалога был заранее прописан, каждое ответвление — предусмотрено разработчиками.

Плюсы подхода

- **Предсказуемость и контроль**
  Точное понимание поведения системы
- **Скорость и дешевизна**
  Минимальные вычислительные ресурсы
- **Простота тестирования**
  Детерминированные результаты

Минусы и болевые точки

- **Хрупкость системы**
  Не понимает синонимы и сложные формулировки клиентов
- **«Деревянный» диалог**
  Клиент сразу чувствует, что говорит с машиной
- **Нулевая гибкость**
  Любое отклонение от скрипта ломает весь диалог

Запрос рынка и иллюзии

Реальность за запросом

Непонимание технологии

Клиенты не видят принципиальной разницы между ChatGPT для общих задач и специализированным промтом для робота в контакт-центре

Ожидание магии

Вера в «магическое» решение без серьезных инвестиций в промт-инженерию, валидацию и постоянную оптимизацию

Недооценка рисков

Игнорирование реальных вызовов: галлюцинации модели, существенная стоимость токенов, непредсказуемость ответов

Наши KPI — Зачем мы пошли на LLM

Главная цель: не заменить, а усилить робота

Мы не стремились полностью отказаться от проверенных подходов, а хотели создать гибридную систему, объединяющую надежность классических решений с гибкостью современных языковых моделей.

Ключевые метрики:

Конверсия

Рост целевых действий: заявки, подтверждения, завершенные транзакции

NPS

Повышение удовлетворенности клиентов от качества диалога

Устойчивость

Способность диалога адаптироваться к вариативности речи клиентов

Провалы

Снижение количества "сломанных" диалогов из-за непонимания

Шаг 1

Правильный промтинг это инженерия

Промт для робота ≠ промт для чата

Создание промтов для роботов КЦ требует системного подхода и глубокого понимания контекста бизнес-процессов. Это не просто «спросить у ChatGPT» — это полноценная инженерная дисциплина.

Решение:

Декомпозиция промта на слои

Системный промт

Пример: *"Ты — вежливый оператор кол-центра компании X. Цель звонка — подтвердить запись клиента на завтра в 15:00. Никогда не придумывай информацию, которой у тебя нет в контексте. Не используй сленг или неформальную лексику. Будь кратким и конкретным."*

роль | контекст | правила игры | строгие табу

Пользовательский промт

Пример: *"Уточни у клиента, действительно ли ему удобно время 15:00, или он хотел бы перенести встречу на другое время. Если клиент хочет перенести, предложи альтернативные слоты: 10:00 или 17:30."*

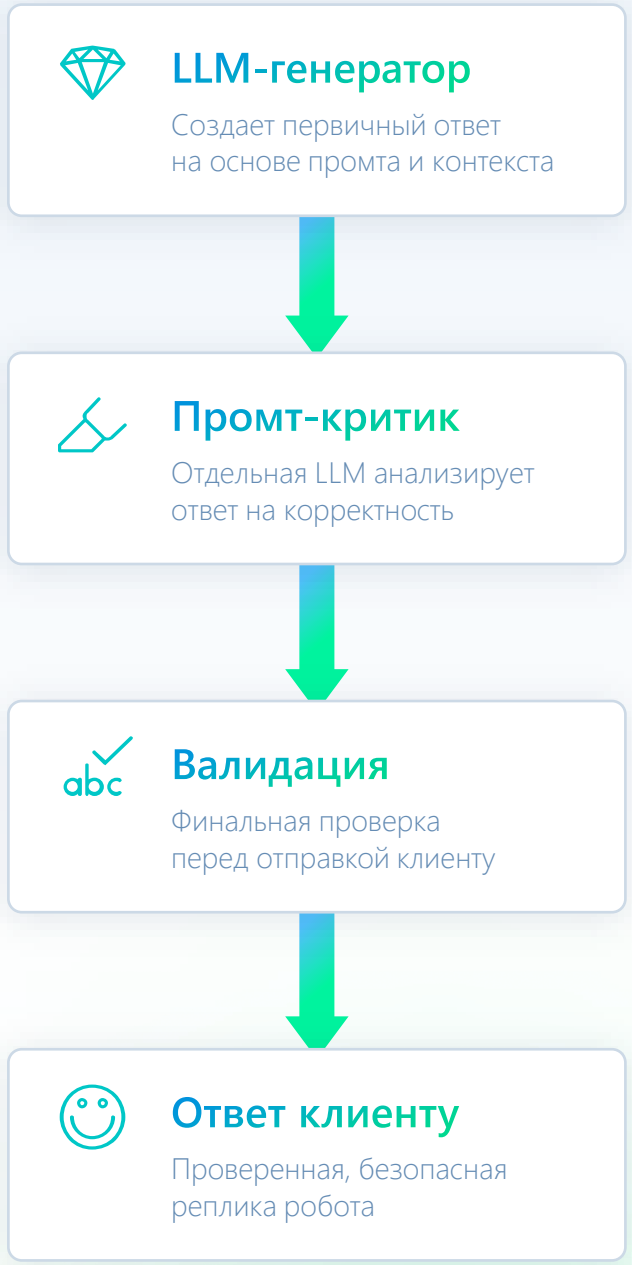
конкретный запрос в рамках текущего диалога

Шаг 2

Борьба с галлюцинациями

Проблема: **LLM может генерировать выдуманные данные**

Языковые модели склонны к «галлюцинациям» — генерации правдоподобно звучащей, но фактически неверной информации. В контакт-центре это критический риск, способный подорвать доверие клиентов.



Критерии проверки промт-критиком

Соответствие фактам

Все данные должны браться только из предоставленного контекста

Отсутствие выдумки

Проверка на галлюцинации и додумывание информации

Соблюдение скрипта

Ответ должен соответствовать бизнес-логике и правилам компании

Шаг 3

Проблема задержек и UX

Критическая проблема:

Процесс валидации и генерации ответа может занимать 10+ секунд. За это время клиент может решить, что линия прервалась, и положить трубку.

Решение:

Заглушки-паузы

Типы заглушек

- «Минуточку, уточняю информацию в системе...»
- «Один момент, загружаю данные...»
- «Сейчас посмотрю детали вашей записи...»
- «Секунду, проверяю доступность времени...»

Результат внедрения

- ↗ Клиент остается на линии и не прерывает звонок
- ↗ Создается впечатление «думающего» робота
- ↗ UX остается на высоком уровне
- ↗ Снижение процента незавершенных диалогов **на 42%**

Шаг 4

Управление бюджетом и фокусом

Проблема: **Бесконечные диалоги сжигают бюджет**

С внедрением LLM клиенты получили возможность вести более естественный диалог. Однако это привело к неожиданному побочному эффекту: некоторые клиенты начали «болтать» с роботом на посторонние темы, сжигая токены и уводя разговор от целевого действия.

Решение:

Механизм ограничения диалога

1 Жесткий лимит ходов

установка максимального количества реплик робота

2 Эскалация на оператора

при невозможности завершить диалог успешно за лимит ходов

3 Приоритизация целевых действий

робот фокусируется на ключевых вопросах, игнорируя отвлекающие темы

4 Мягкое подведение к цели

после N-го шага робот начинает деликатно возвращать диалог к основной задаче

Достигнутый эффект

- ↗ Контроль стоимости звонка
- ↗ Сокращение средней длительности звонка **на 23%**
- ↗ Фокусировка на целевом результате
- ↗ Повышение конверсии в целевое действие

Парадигма тестирования: смена подхода

Было: эра REGEX

Скрипт ▸ **Детерминированный результат**

Можно построить полную карту переходов между состояниями. Каждый вход гарантированно дает один и тот же выход. Тестирование предсказуемо и однозначно.

- 100% воспроизводимость результатов
- Четкие граничные условия
- Детерминированные пути выполнения

Стало: эра LLM

Промт ▸ **Вероятностный результат**

Невозможно предсказать точную формулировку ответа. Даже идентичный промт может давать разные, хотя и семантически близкие результаты. Требуется новая философия тестирования.

- Оценка семантической корректности
- Статистический подход к качеству
- Тестирование на вариативность входов

Этот фундаментальный сдвиг требует переосмысления всех процессов контроля качества и создания принципиально новых инструментов тестирования AI-систем.

Как мы тестируем LLM-робота

Мы разработали комплексную трехуровневую систему тестирования, которая позволяет оценивать качество робота с разных углов и обеспечивает уверенность в его надежности.

Сценарное (регрессионное) тестирование

Обработка рабочих диалогов
(с Regex-эпохи) через новую LLM-систему

Цель: **Робот не стал работать хуже, и сохранил уровень качества**

Тестирование на вариативность

Один запрос задается десятками
разных способов — используя синонимы,
жаргон, сложные грамматические конструкции

Цель: **Выросла устойчивость диалога к естественной вариативности человеческой речи**

Инструменты для тестирования

Автотесты

Автоматизированные скрипты, проверяют ответы LLM на все необходимые ключевые элементы и отсутствие запрещенной информации

Речевая аналитика

AI-агенты по ночам автоматически прослушивают и анализируют все диалоги за день, находят отклонения от скрипта, некорректные формулировки и потенциально проблемные ситуации

Системы мониторинга

Отслеживание метрик в реальном времени: средняя длительность звонка, успешно завершенные диалоги, потраченные токены, NPS, эскалации на оператора

Помимо методологии, мы создали технологический стек инструментов, которые автоматизируют процесс контроля качества и позволяют отслеживать проблемы в режиме реального времени.

Выводы и итоги

1

LLM — мощный инструмент, но не панацея

Это технология, которую необходимо грамотно интегрировать в существующие бизнес-процессы, а не слепо применять везде

2

Ключ к успеху — системный подход

Не в «волшебном» промте дело, а в комплексной системе: промт-инженерия + многоуровневая валидация + контроль диалога + адаптивное тестирование

3

Гибридная архитектура побеждает

Мы создали систему, которая сочетает гибкость и естественность LLM с надежностью и предсказуемостью классических инженерных подходов

Переход на LLM — это не революция, а эволюция. Успех приходит к тем, кто понимает ограничения технологии и строит вокруг нее правильную инженерную инфраструктуру.

Вызовы

Внедрение агента-коллектора на базе LLM

— полностью автономный
— незаскриптованный
— «креативный»
— на базе RAG

Динамические правки скриптов

более 300 доработок с начала года.
Качество ~90%

Умные автоответчики

95% диалогов с АО завершаются автоматом
благодаря AMD моделям, 5% — нет

Сложность покрывать все звонки фильтрами

0,7–1% звонков *(20 тыс. в день)*
«отлавливаются» речевой аналитикой
как потенциальные отклонения

Решение

Переход к agentic-парадигме QA
с GigaChat Max 2 под капотом

Мы комбинируем существующие инструменты PA с новыми возможностями LLM

В результате получается «конвейер» для массовой работы с коммуникациями в режиме 24/7 и 100% интеллектуальным покрытием

Снижаем выборку до нескольких тысяч

LLM – это не бесплатно

Предварительный отбор

Запрос: «реструктуризация, контакт»

Анализируем

Промт

«проверка и улучшение агента»

Грамотно используем ресурсы

Планировщик

Работаем 24/7

— **Днём** активный обзвон

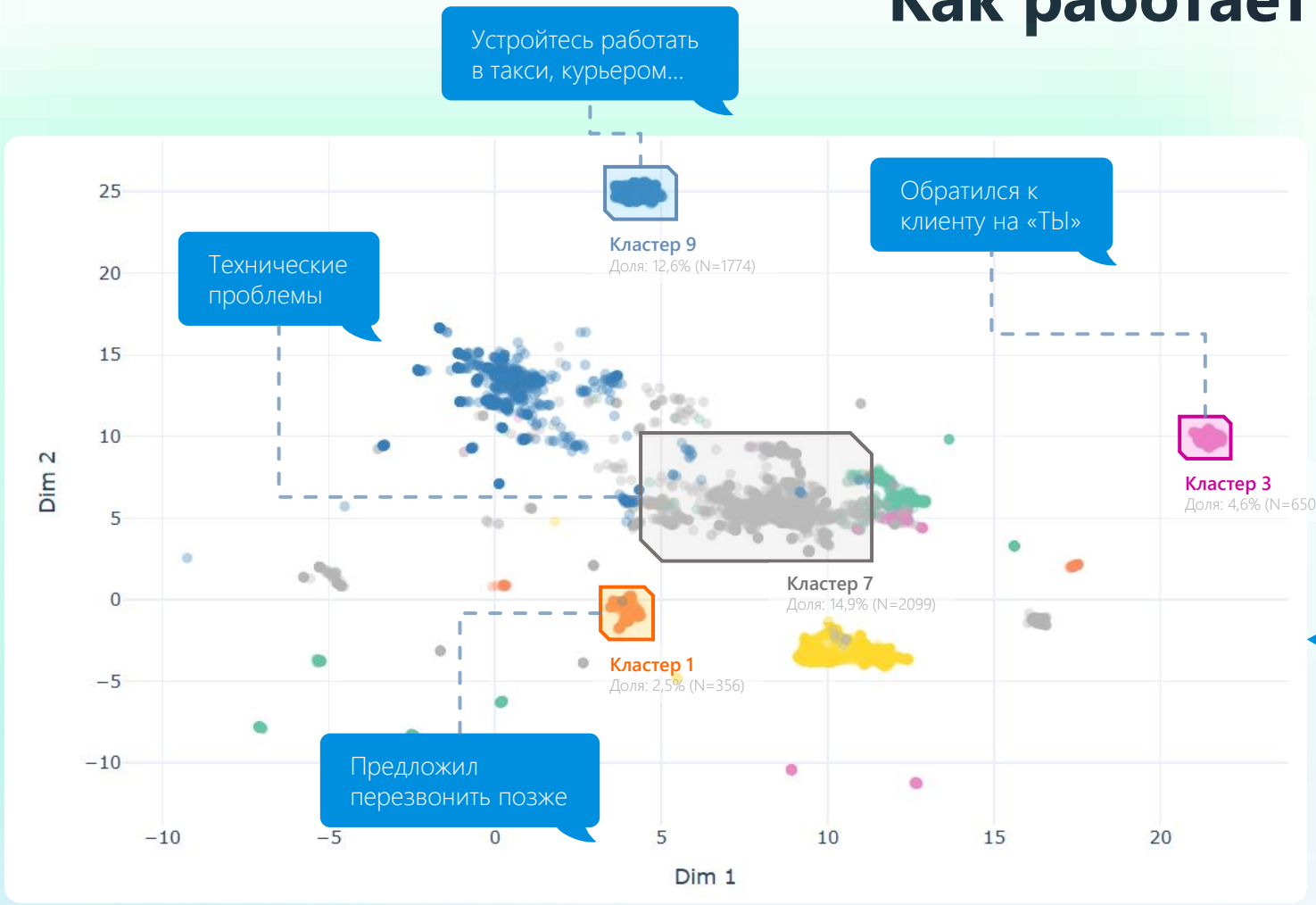
— Все ресурсы брошены на робота

— **Ночью** «разбор полетов» и аналитика

— GigaChat трудится 24/7

Концепция «AI-закрытого цикла»

Как работает аналитика и улучшения



ЭТАП ГРУППИРОВКИ

Мы работаем с проблемной выборкой, чтобы выделить основные группы проблем с помощью AI

87%+ звонков прошло корректно

2% отсеиваются

11% выборка для улучшений

1 **Результат AI-аналитики**

2 Анализируем отклонения через AI

3 Правим исходного агента на основе AI

Как работает аналитика и улучшения

- AI формирует рекомендации внутри кластера
- Близкие и частотные по значению рекомендации объединяются

ЭТАП АНАЛИТИКИ И РЕКОМЕНДАЦИЙ

Внутри групп проблем анализируется каждый звонок в несколько этапов для генерации рекомендаций

Рекомендация: внедрить чек-лист перед генерацией ответа

AI генерирует улучшения исходных промтов

Перед отправкой ответа проверь:

- [ ] **Я не предложил перезвонить через несколько часов**
- [ ] Я не придумал суммы/сроки — информация только из контекста.
- [ ] **Я не обращаюсь к клиенту на «ты».**
- [ ] **Я не предложил найти работу в такси или курьером**
- [ ] Если клиент просил живого оператора/связь плохая — я эскалирую.

Если любой пункт нарушен — исправь реплику и только потом отправляй.

1 Результат AI-аналитики

2 Анализируем отклонения через AI

3 Правим исходного агента на основе AI

Концепция «AI-закрытого цикла»

AI-агент коллектор общается с клиентом

Робот общается с клиентом по скрипту

Живой оператор общается с клиентом

AI-аналитик анализирует все звонки, оценивает корректность и подсвечивает проблемы

AI выделяет основные проблемы и улучшает исходные промты **или дает рекомендации**

Нам удалось добиться впечатляющих результатов

Внедрение AI позитивно повлияло на бизнес метрики и позволило решить ранее «неподъемные» задачи

с 90% до 98%

повышено качество диалогов

День-в-день

скорость выявления новых автоответчиков

Обещания

+3,1%

Сбросы

–3,1%

Нераспознанных АО

1% *(–3,7%)*

на 73%

сокращена длительность диалогов с автоответчиками

6 600 человек

потребовалось бы для достижения такого покрытия

300+

доработок скриптов робота реализовано через AI-анализатор багов