kvint | voice assistant for business

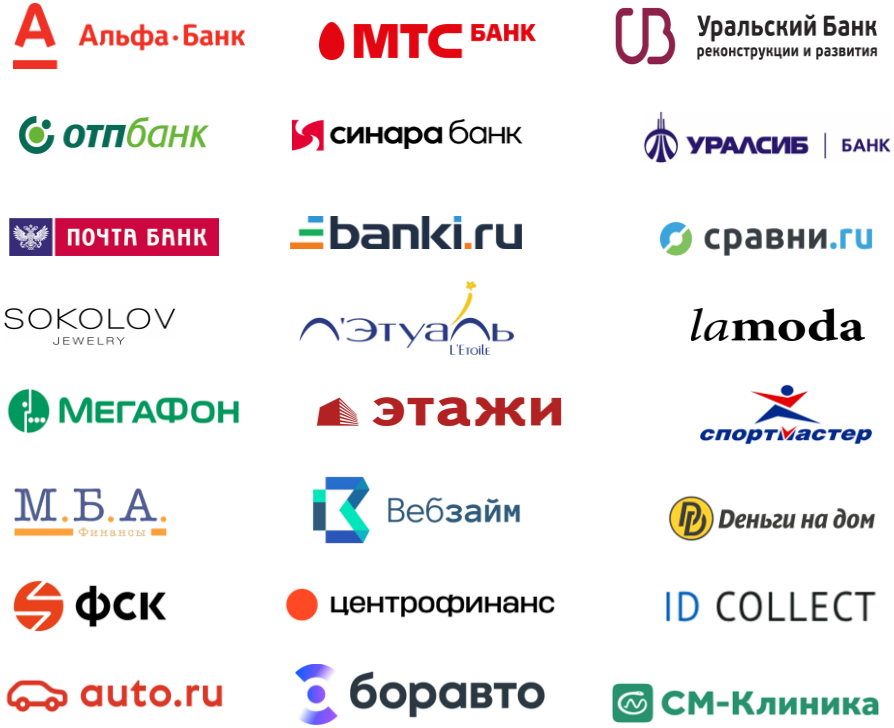
Как КВИНТ создает генеративного голосового робота для входящей линии

СПИКЕР:

Захарищев Михаил
Руководитель направления развития бизнеса

m.zak@kvint.io
+7 922 931 1888

Клиенты и пилотные проекты



ТОП 7 Сегментов

- Банки
- Телеком
- Ритейл
- Медицинские учреждения
- Коллекторские компании
- Недвижимость
- Транспортные компании

Компания в цифрах

400+ Внедрено более 400 роботов

90+ Команда разработки роботов

6+ Более 6 лет на рынке

2000+ Более 2000 звонков в минуту

5 млн 5 млн звонков в день

Путь от скрипта к гибриду и результаты

kvint | voice assistant for business

Эволюция голосовых роботов КВИНТ

Скрипт

Функционал робота ограничен рамками скрипта

Гибрид

LLM ускоряют часть процессов в архитектуре робота, но напрямую с абонентом не общаются

Генеративка

Ответы робота генерируются моделью. Реализация полностью на потоковом синтезе

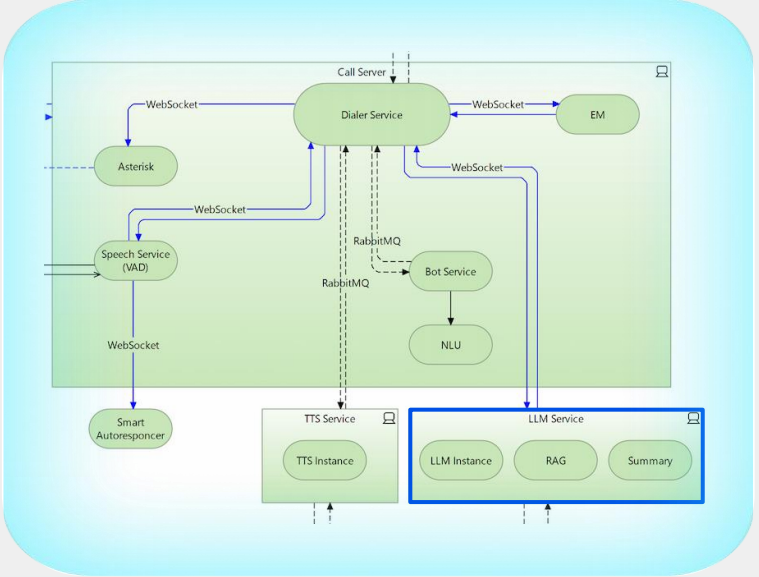
2018-2024 → 2023-2024 → 2024

Отличие гибридного робота от скриптового

В основе диалога – прописанный скрипт.

Функции LLM:

- Создание дата-сета для NLU модели
- Генерация кейсов для тестирования
- Усовершенствование определения намерения
- Составление саммари по диалогу



Результаты перехода на гибридных роботов в 2023

+10-15%

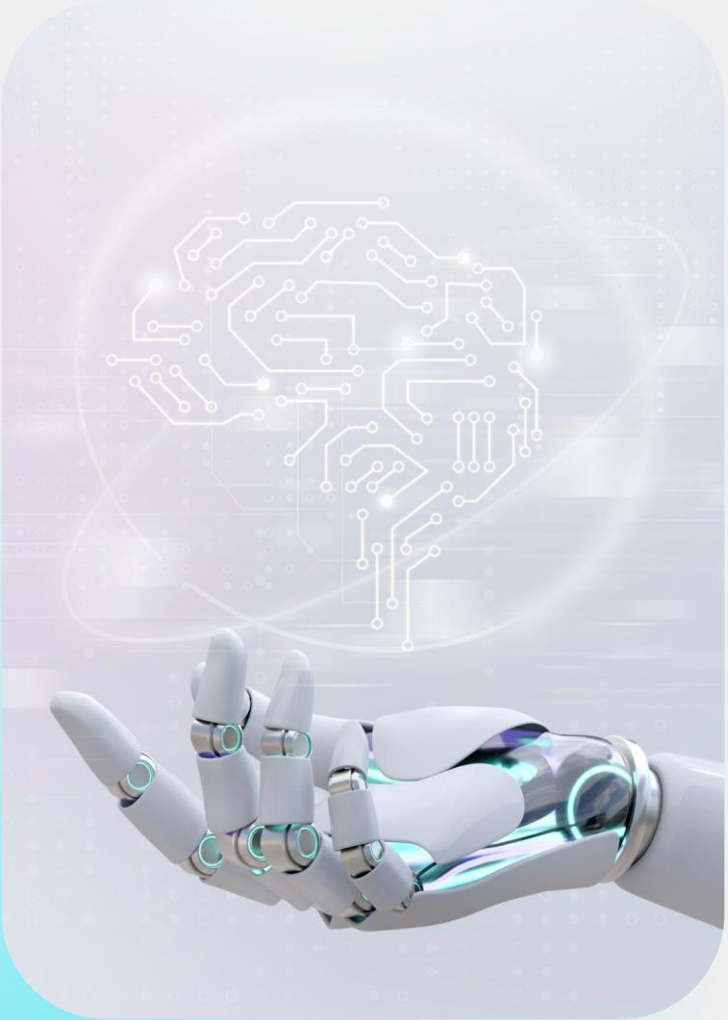
Рост целевой конверсии

+20% Более быстрое создание скрипта

+13% К точности робота

+0₽ Рост стоимости разработки

Переход на генеративных роботов



Возможности решения

Замена IVR на умного робота

Взаимодействие с пользователем без громоздких многоуровневых меню. Абонент задает вопрос естественным языком и быстро получает четкий ответ.

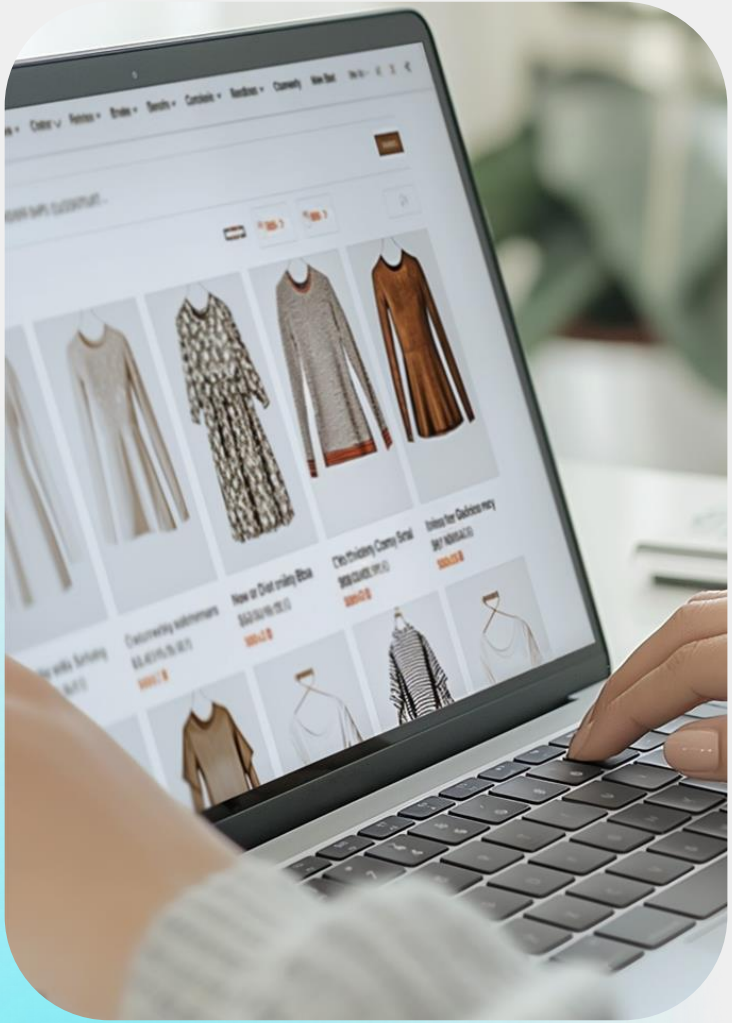
Взаимодействие с клиентами становится более персонализированным и эффективным

Возможности решения

Ответы на входящие запросы

Робот способен обрабатывать входящие звонки, обеспечивая мгновенные ответы на запросы клиентов, распознает и интерпретирует широкий спектр вопросов, от простых до более сложных.

Снижение нагрузки на операторов колл-центра и ускорение процесса обслуживания



Возможности решения

Консультирование по продуктам и услугам

Может отвечать на вопросы о характеристиках товаров, условиях покупки, акциях и других аспектах, связанных с предложениями компании.

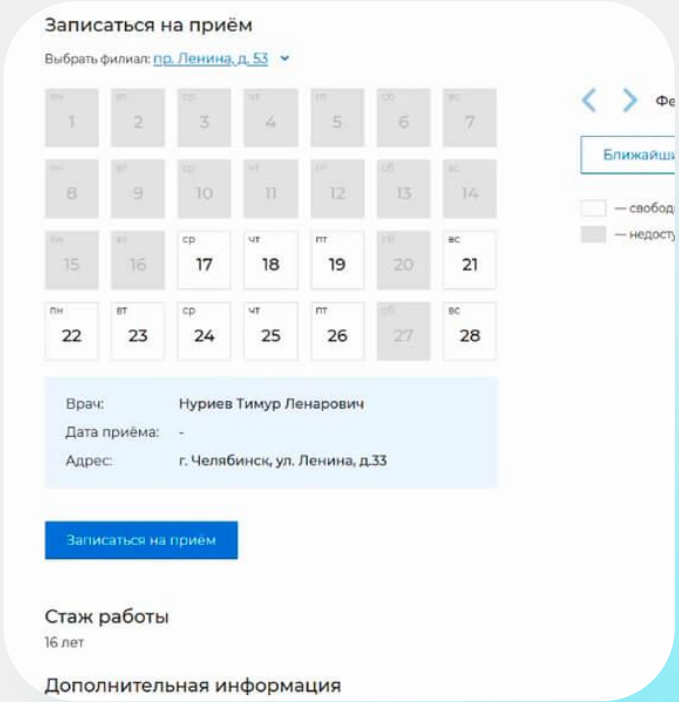
Повышение удовлетворенности и лояльности клиентов

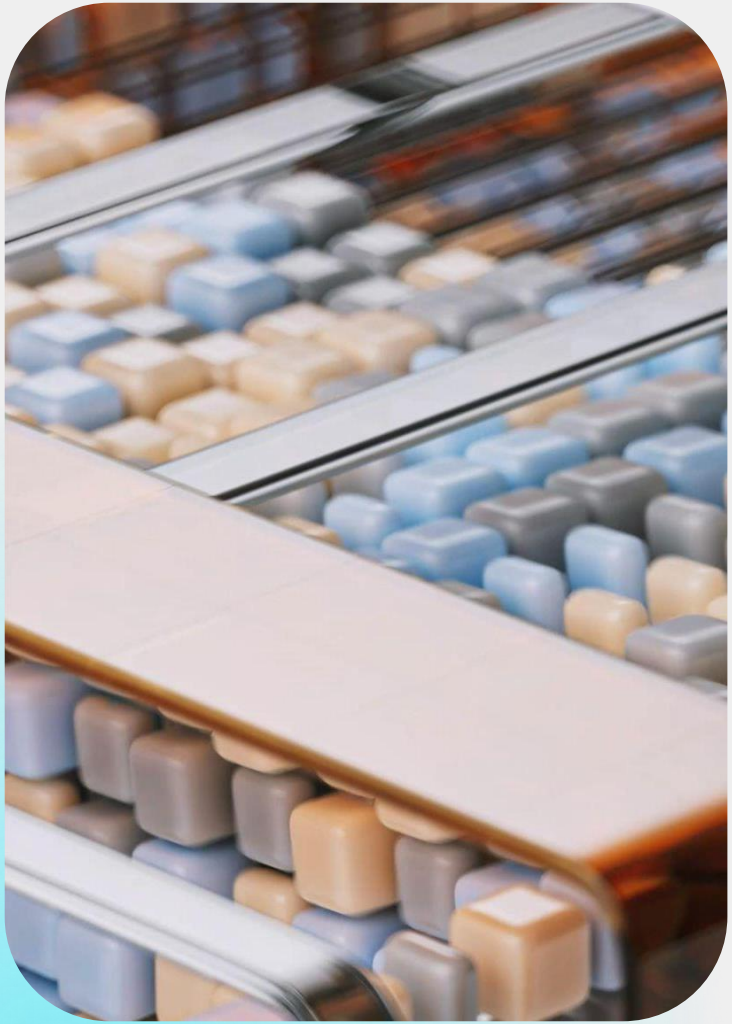
Возможности решения

Онлайн запись, бронирование

Голосовой робот может управлять календарем, проверять доступные временные слоты и фиксировать бронирования в соответствии с предпочтениями клиента.

Минимизирование вероятности ошибок при вводе данных.





Возможности решения

Интеграция с внутренними системами Клиента

Интегрируется с внутренними системами Клиента. Это обеспечивает доступ к актуальной информации о пользователях, заказах и других данных, необходимых для качественного обслуживания.

- Персонализированные ответы и рекомендации на основе истории взаимодействий и предпочтений клиента

Преимущества решения

СКОРОСТЬ
Ускоренное время создания робота

РАСШИРЕНИЕ ФУНКЦИОНАЛА
Создание больших разветвленных роботов

ПЕРСОНАЛИЗАЦИЯ
Генерация разных ответов

ГИБКОСТЬ
Обучение на базе знаний Клиента

Для каких компаний подходит

- Консультант на входящей линии для крупного и среднего бизнеса
- Создание онлайн-приемной в государственных и медицинских учреждениях
- Реализация бронирования в сетях ресторанов, салонах красоты, автосервисах

Особенности разработки

Облако и контур: какую модель выбрать

Контур

- → Большая конфиденциальность данных
- → Подходит для крупных компаний

- Реализация только на российских LLM (Vikhr, YandexGPT или Giga Chat)

Облако

- → Не предполагает работы с конфиденциальными данными пользователей (счета, задолженности и т.д.)

- ChatGPT
- Российские аналоги

Работа с генеративными моделями

С какими сложностями мы столкнулись в рамках тестирования LLM и как смогли решить проблему

Проблема	Как решали	Результат
Точность ответов по базе знаний	Настройка параметров моделей	Зависит от запроса клиента. Модель может как просить переформулировать вопрос, если не знает ответ, так и генерить “креативный” ответ на основе базы знаний
Поиск по базе знаний	Использование векторных моделей	Поиск нужного контекста для ответа за 0.2 секунды
Низкая скорость генерации ответа	Подбор параметров генерации. Параллельная работа LLM и потокового синтеза	Генерация ответа за 2-4 секунды
Необходимость доработки LLM для контура	Использование few-shot prompting	Задаем поведение модели для ряда самых распространенных диалогов. Тогда она точно знает что отвечать на самые частые вопросы

Работа с потоковым синтезом

Особенности

- → Основной критерий выбора – комбинация скорости синтеза и качества.
- → Вручную прослушивать тестовые звонки для выбора наиболее подходящего решений
- → Тестировали синтезы Яндекса, Тинькова и Сбера
- → Генерация на русском языке только предложениями

Задержки потокового синтеза

0.3-0.5 секунд

Используем прием, где **задержка** от вопроса пользователя до старта ответа робота **составляет до 1.5 секунд**

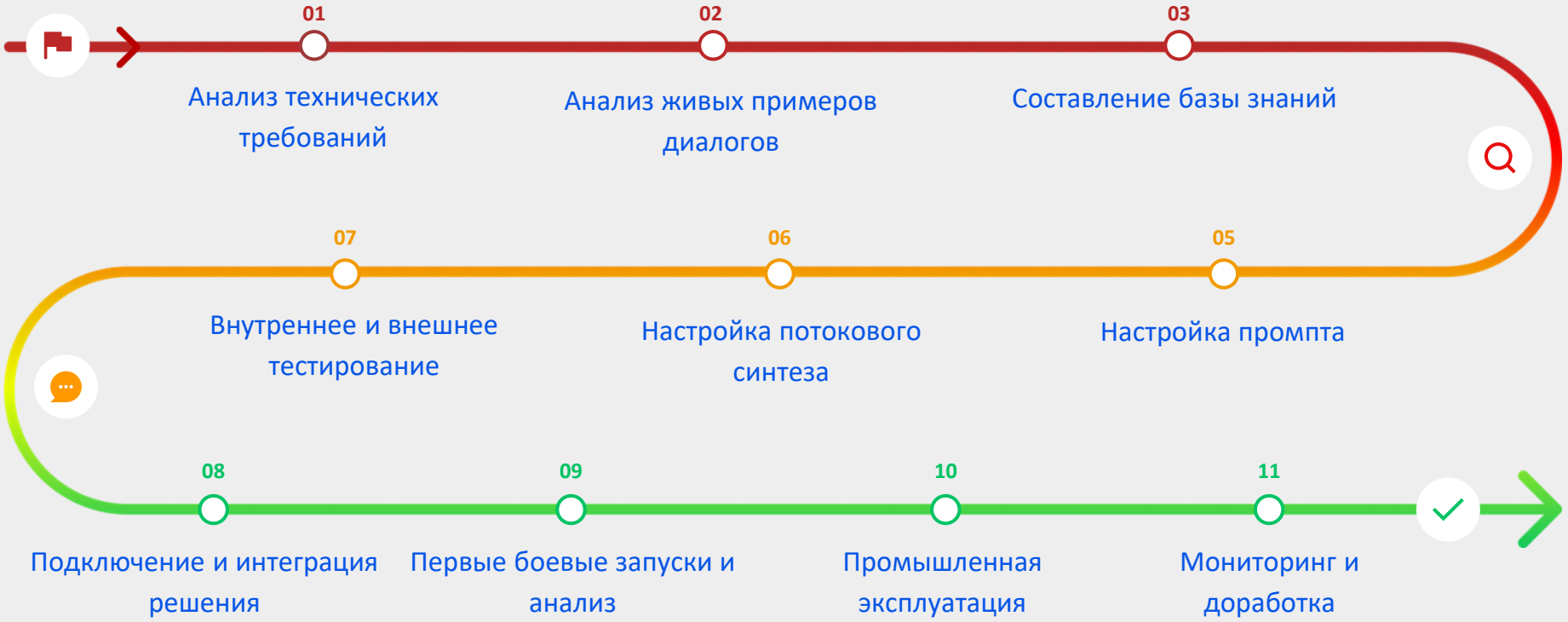
Внедрение

Запуск первого пилотного проектов

- **Октябрь 2024** – первый демо-робот на русском языке
- Решение полностью “под ключ”. Мы сами создаем робота и дорабатываем его
- Полное сопровождение проекта и оперативные ответы от команды КВИНТ

Послушать пример диалога

Как выглядит процесс внедрения



Интеграции с какими платформами и сервисами мы предлагаем

- CRM
- Системы бронирования
- АТС
- HR-сервисы
- Мессенджеры

kvint | voice assistant for business

Подходите на стенд для демонстрации функционала роботов КВИНТ

Захарищев Михаил

Руководитель отдела развития бизнеса

+7 922 931 1888
m.zak@kvint.io
www.kvint.io