

ИИ-оператор, которого не просят переключить на человека

Как добиться самостоятельного решения
клиентских задач

Голос проверяет ИИ на прочность

В телефонном разговоре ИИ-оператор должен понимать клиента, принимать решения и реагировать **в ритме естественного диалога**

Любая задержка или ошибка становится частью клиентского опыта

Текстовые каналы



Телефонный канал



Темп коммуникации

Клиент готов подождать ответ

Продолжительная пауза воспринимается как зависание или потеря контекста

Понимание клиента

Система получает и обрабатывает готовый текст

Система должна распознать речь, отделить шум, обработать паузы и перебивания

Что оценивает клиент

Смысл ответа и результат

Содержание, скорость реакции, голос и поведение в разговоре

Что значит «задача клиента решена»?

Показатель автоматизации отвечает на вопрос, **кто провёл разговор**, но имеет ограничения:

01



Не показывает результат
для клиента

02



Не учитывает качество
клиентского опыта

03



Не ставит целью решить
задачу сразу

Чтобы оценить ИИ-оператора, нужно понять, **достиг ли клиент цели обращения.**

Сколько задач ИИ-оператор решает сам?

TSR

Task Success Rate

как ключевая метрика
ИИ-оператора



Доля клиентских задач, которые
ИИ-оператор самостоятельно довел
до успешного предусмотренного результата
с соблюдением требований к качеству и времени

Единица оценки

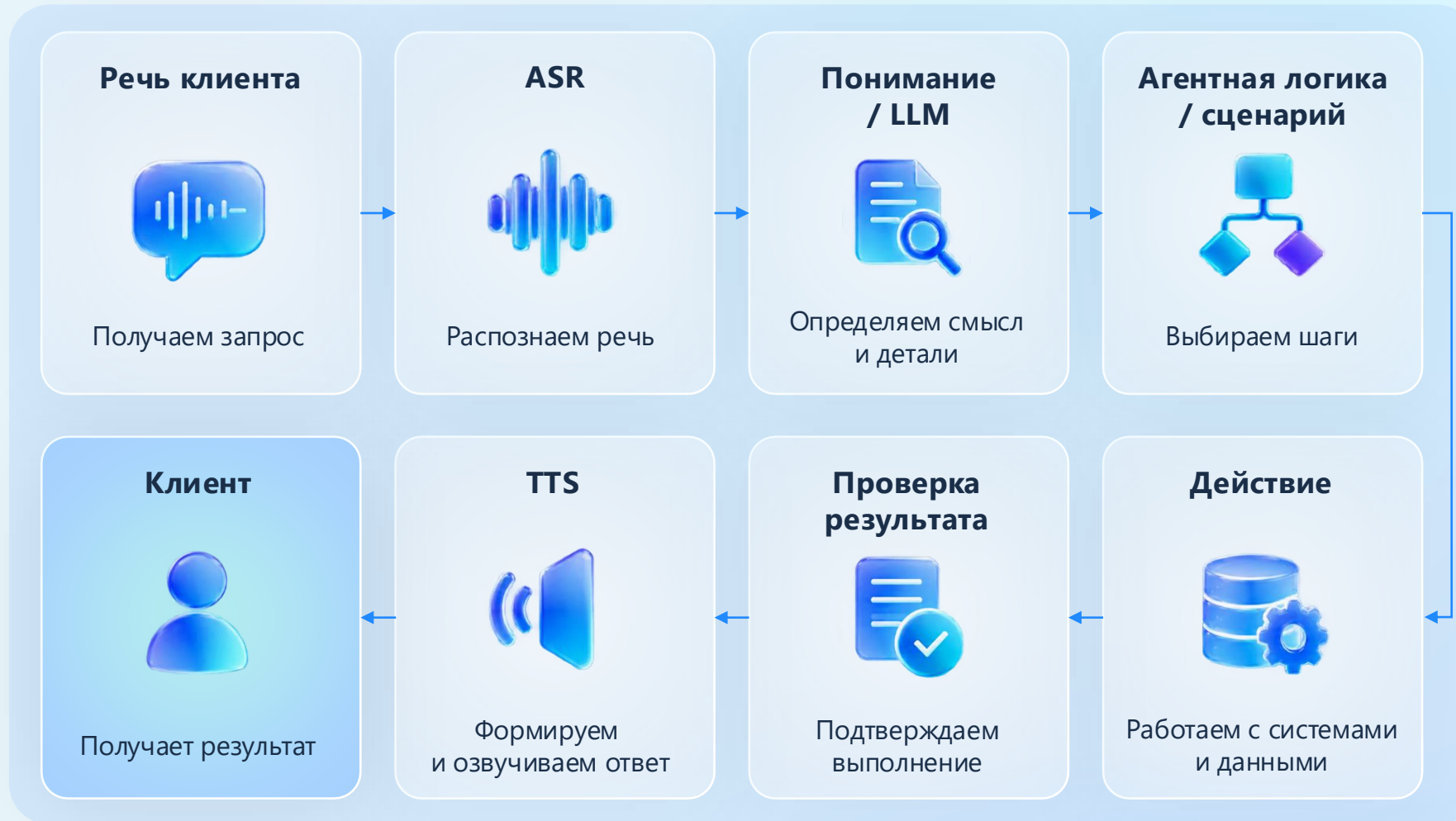
Клиентская задача целиком
– от запроса до результата

Показатель всей системы

Итог совместной работы технологий
понимания речи, выполнения действий
и голосового взаимодействия

ИИ-оператор

как технологическая цепочка



Ошибка любого звена может прервать выполнение задачи до результата

Усиливаем весь технологический контур, в котором каждая составляющая влияет на итог клиентской задачи

Сквозной слой: latency и управление диалогом

Звонок автоматизирован, но задача не решена

01 Распознавание речи

Пауза принята за конец речи

*Запишите меня... эээ...
к кардиологу... восьмого
июня после...*

ИИ-оператор начинает
обработку, не дождавись
смыслового завершения фразы.
Запрос остается неполным.

02 Latency

Задержка прервала задачу

*Да, запишите
на 18:30.*

Во время обращения к системе
ИИ-оператор молчит. Клиент решает,
что связь потеряна, и завершает
звонок до подтверждения

03 Перебивание и состояние диалога

Корректировка не изменила действие

ИИ-оператор

Записываю вас на среду

Клиент

Не в среду, а в четверг

ИИ-оператор продолжает
исходный сценарий и выполняет
действие — создает запись
на неправильную дату

04 LLM и агентная логика

Нарушен порядок действий

*Перенесите на 8 июня после
18:00. Если нет свободных
мест, оставьте как есть*

ИИ-оператор отменяет
текущую запись до
проверки нового слота.
Клиент теряет запись.

Качество — больше, чем правильный ответ

В голосовом канале клиент одновременно воспринимает:

Что сказал ИИ-оператор

01

Правильно понял
и ответил на вопрос
клиента

Как ИИ-оператор это сказал

02

Естественно звучит
голос, интонация,
темп, паузы

Как ИИ-оператор ведет себя в разговоре

03

Вовремя отвечает
и не зависает, реагирует
на перебивания, не говорит
параллельно с клиентом

Качество — больше, чем правильный ответ

Анализируем не только, **что говорит**,
но и **как говорит** ИИ-оператор



Исследования подтверждают:

уверенная утвердительная интонация и относительно быстрый темп речи способствуют **снижению недовольства клиентов и вероятности повторного обращения** в течение 24 часов¹.

[1] – Источник: Zhou et al., 2025, Journal of Business Research. Serve with voice: The role of agents' vocal cues in the call center service

Клиент должен хотеть продолжать разговор



Естественные голос и
отклик на речь клиента,
быстрая реакция



Меньше раздражения,
преждевременных
завершений и переводов
на сотрудника



Клиент остается
в диалоге до результата

В нашей модели это становится частью

Task Success Rate

Если клиент не хочет продолжать разговор с ИИ-оператором,
тот физически не сможет довести задачу до результата.



От дикторского голоса к живой речи

Четыре стиля моделей генеративного синтеза **SaluteSpeech** под разные сценарии

Freespeech

Спонтанная разговорная речь — для свободного живого диалога

Lifelike

Естественность и актерская выразительность

Operator

Манера оператора КЦ: внимание, уверенность, уместная подача

Dictor

Четкая нейтральная подача — для формальных ситуаций и информационных сообщений

Выбор зависит от сценария и роли ИИ-оператора

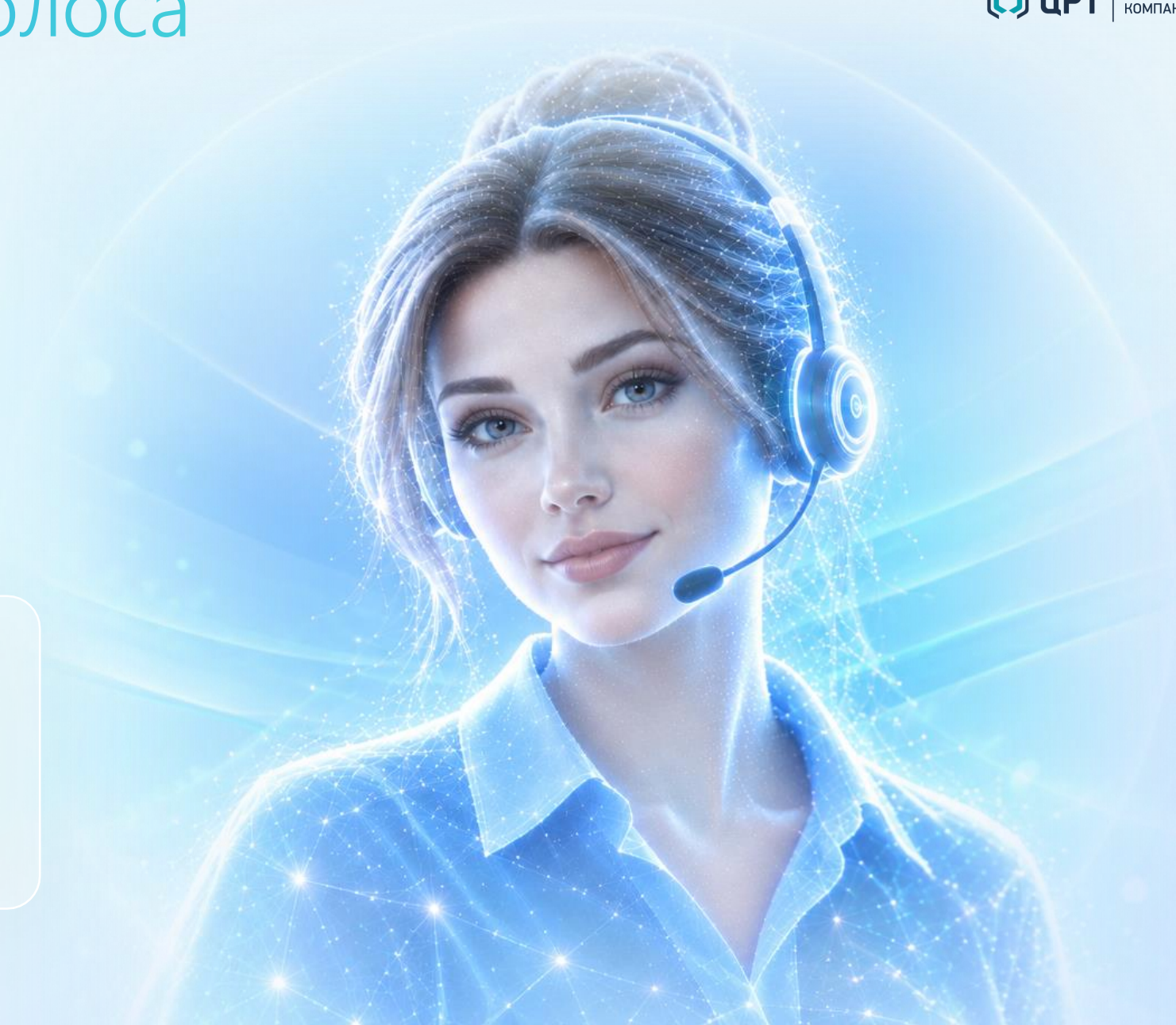
Голос ИИ-оператора не гарантирует решение задачи,
но естественная речь способствует продолжению диалога

От дикторского голоса к живой речи



Operator

Манера оператора КЦ: внимание, уверенность, уместная подача



Что показывают слепые тесты

Freespeech

Спонтанная разговорная речь

Lifelike

Актерская выразительность

Operator

Манера оператора КЦ

Dictor

Четкая формальная

GigaTTS vs прошлое поколение TTS



Доля предпочтений, %
Слева – первая модель в паре

Lifelike vs Dictor



SBS

Какое аудио звучит лучше

HVR

Какое звучит естественнее

Lifelike vs Freespeech



По естественности и живости звучания:
Dictor < Lifelike < Operator ≈ Freespeech

86,5%

pSER — доля фраз без дефектов произношения

От очереди реплик к живому ритму разговора

Клиент думает и формулирует ответ

Длинная фраза с паузами, «эээ», подбором слов; диктовка данных по частям

Клиент вмешивается в речь ИИ-оператора

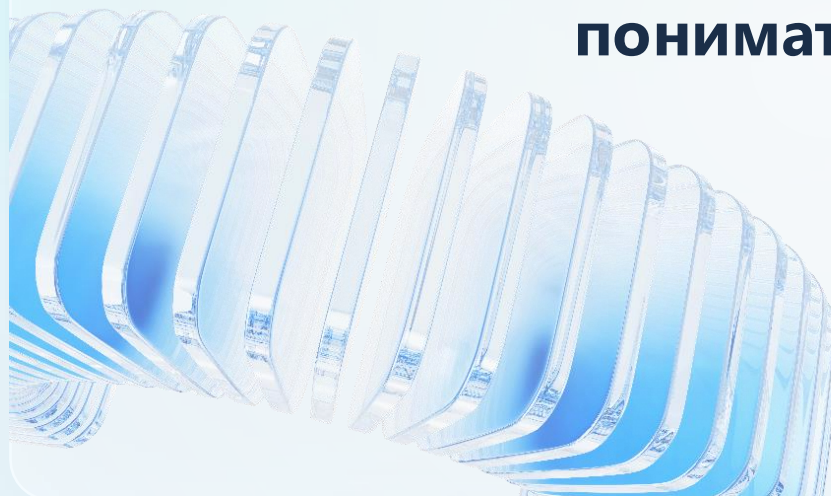
«Да», «нет», «ага», «угу», «стоп», исправление данных, встречный вопрос

Клиент говорит одновременно с ИИ-оператором

Одновременная речь; фоновые разговоры/шум

Правильный ответ в неправильный момент — это плохой диалог.

Голосовой канал требует **понимать динамику разговора.**



От очереди реплик к живому ритму разговора

Клиент думает и формулирует ответ
Длинная фраза с паузами, «эээ», подбором слов; диктовка данных по частям

✧ Не перебивать раньше времени и не ждать слишком долго

Speaker Turn модель + VAD + потоковое распознавание и синтез

Клиент вмешивается в речь ИИ-оператора
«Да», «нет», «ага», «угу», «стоп», исправление данных, встречный вопрос

✧ Мгновенно отличать содержательную реплику от той, которую можно пропустить

Speaker Turn → распознавание → классификация → принятие решения без потери latency

Клиент говорит одновременно с ИИ-оператором
Одновременная речь; фоновые разговоры/шум

✧ Различать реальное перебивание, фоновый шум и параллельную речь

Гибкое управление синтезом: приостановка и возобновление или новый ответ в зависимости от поведения клиента

Между вопросом клиента и ответом ИИ — гонка за миллисекунды

Оптимизируем время от окончания речи клиента до первого звука ответа



Свободный диалог, управляемое выполнение

Клиент

*Перенесите доставку на пятницу после шести...
лучше в офис, а не домой*

LLM: свобода диалога

Гибкость в понимании клиента и его формулировок

- Определение смыслов и извлечение данных
- Удержание контекста и понимание исправлений
- Предложение уточнений и формулирование ответа

ИИ-оператор

Уточните адрес вашего офиса

Правила и интеграции: выполнение

Управление данными, развилками и действиями

- Валидация данных и выбор допустимого действия
- Проверка условий и вызов разрешенных функций
- Проверка результата в целевой системе

ИИ-оператор

Давайте проверим ваш заказ...

Как LLM влияет на Task Success Rate

Сохраняем условия клиента

01

LLM понимает уточнения.
Сценарий их фиксирует
и учитывает при выполнении

✧ **Снижаем риск**
выполнить не ту задачу

Сокращаем задержки

02

Меньше дополнительных
обращений к модели
для формальных проверок

✧ **Снижаем риск**
потерять клиента до результата

Проверяем контрольные точки

03

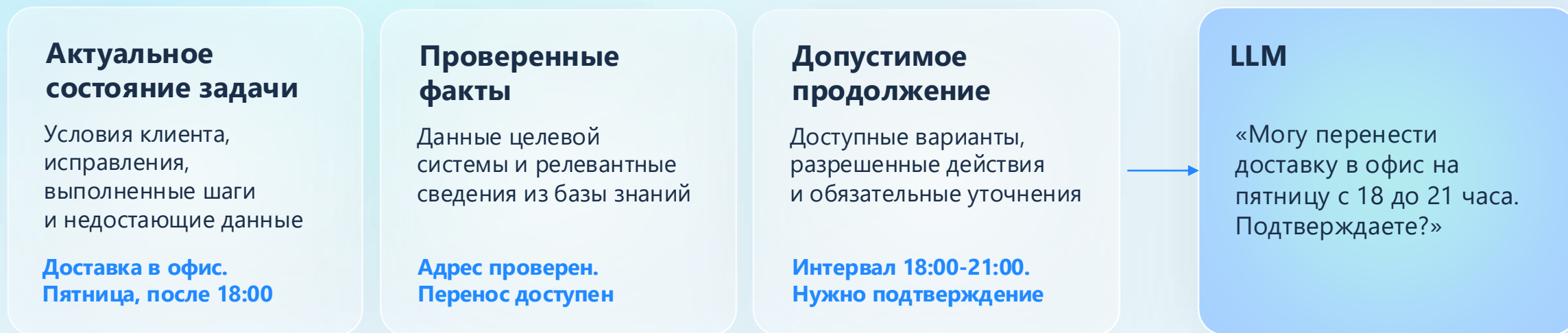
До действия проверяем
условия. После действия
проверяем результат в системе

✧ **Снижаем риск**
объявить успех без результата

Как управлять свободой ИИ

В телефонном разговоре допустимая задержка ограничивает многоступенчатую проверку LLM. Значительную часть контроля закладываем до ее вызова.

Что получает LLM на текущем шаге



✧ Ограничения в промпте ✧ Управляемый контекст

Общие запреты и требования к поведению LLM

Не предлагай недоступное время доставки

Факты, результаты проверок и разрешенные действия для текущего шага

Передаем интервалы, проверенные для данного адреса

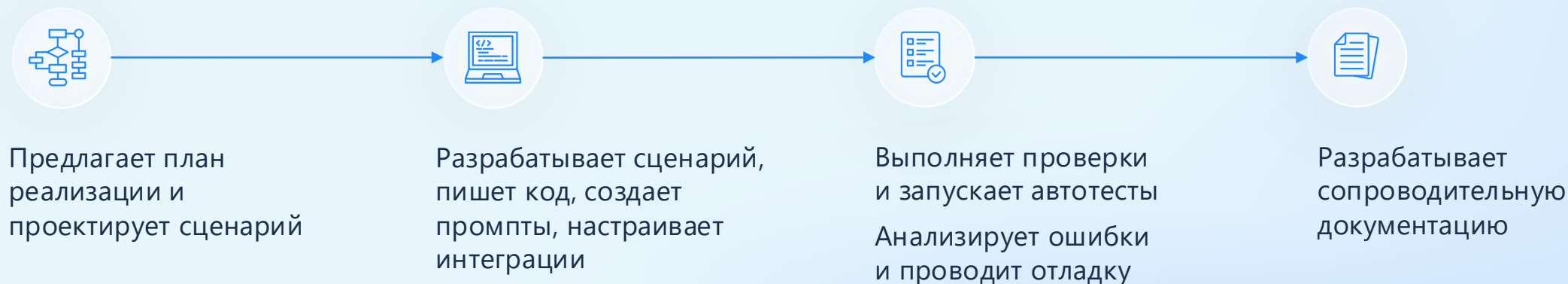
Как быстро создавать сценарии для ИИ-оператора

Ускоряем работу с помощью ИИ-агента для разработки сценариев ИИ-оператора

Принимает на вход требования, спецификации, описания ошибок, предложения нового функционала или продуктовые гипотезы

Разворачивается **вместе с продуктовой тестовой средой** для быстрого запуска, тестирования и отладки сценариев

Как действует ИИ-агент



Как технологическая цепочка работает на Task Success Rate

Звено цепочки

Понимание / LLM

Агентная логика / сценарий

Действия / интеграции

Проверка результата

TTS

Сквозной слой: latency и управление диалогом

Влияние на Task Success Rate

✧ Меньше неверных ответов и действий

✧ Меньше пропущенных шагов и нарушений ограничений

✧ Ниже риск частичного выполнения, необработанного отказа или отсутствия фактического результата

✧ Исключает ложный успех, дает основание для расчета TSR

✧ Ниже риск ухода или запроса оператора до результата

✧ **Меньше потерянных уточнений, срывов задачи и уходов клиента**

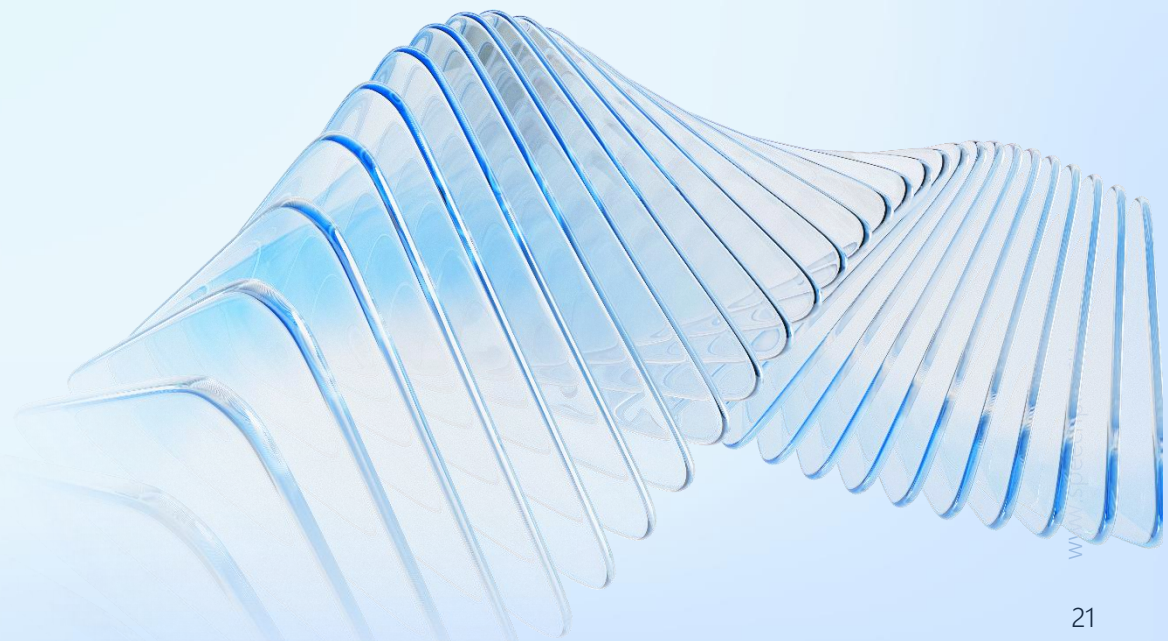
Как мы измеряем Task Success Rate

Какие задачи оцениваем

- **Консультация**
По справочнику или базе знаний
- **Действие в системе**
Создание записи или заявки, отмена, фиксация обращения или жалобы, предоставление сведений из системы, изменение данных в системе
- **Информирование с подтверждением**
- **Предусмотренная маршрутизация**
На подходящего оператора / скилл-группу
- **Завершенный опрос или анкета**

$$TSR = \frac{\text{Число задач с успешным результатом}}{\text{Общее число целевых задач в выборке}} \times 100$$

1 диалог ≠ 1 задача



Как мы измеряем Task Success Rate

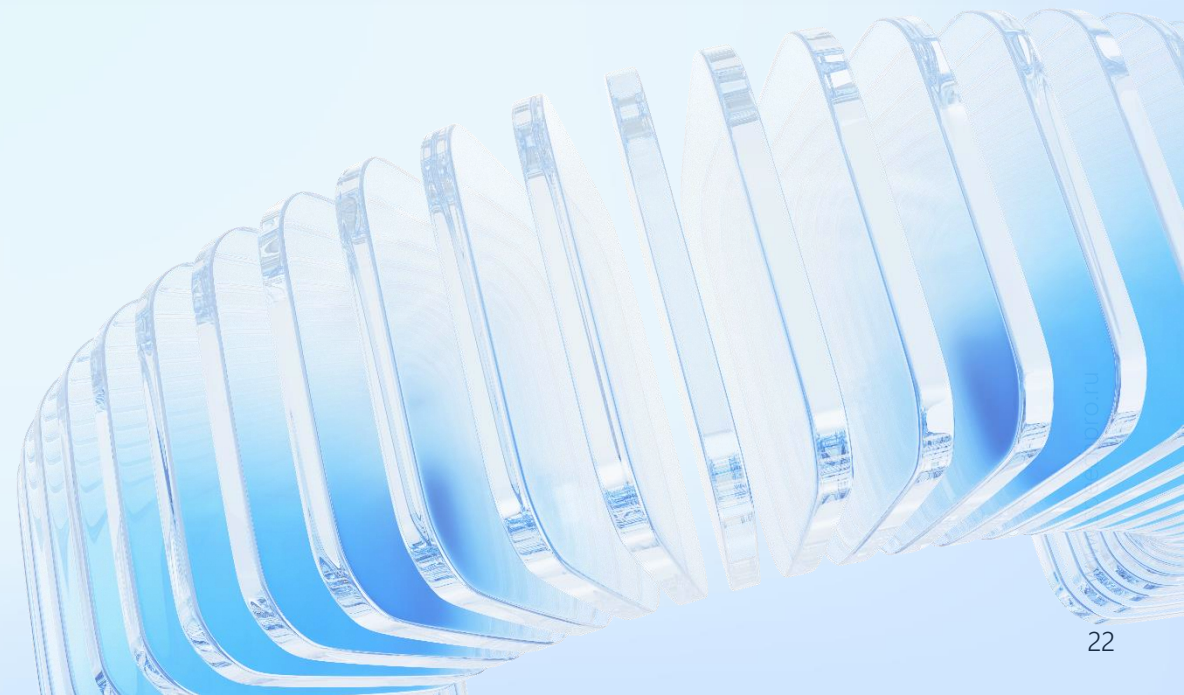
01 **Фиксируем задачи и критерии успеха**

02 **Формируем выборку**

03 **Выделяем задачи внутри диалогов**

04 **Проверяем результат каждой задачи**

$$TSR = \frac{\text{Число задач с успешным результатом}}{\text{Общее число целевых задач в выборке}} \times 100$$



От показателя TSR к причинам потерь

Что анализируем для нерешенных задач

- ✗ Неверный, отсутствующий или частичный результат
- ✗ Незапланированный перевод на оператора
- ✗ Уход клиента до получения результата

Локальные метрики

Понимание запроса

Полнота действий

Соблюдение правил

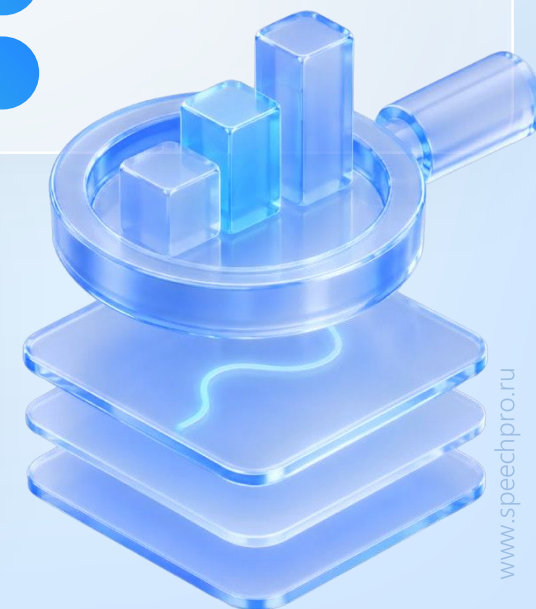
Корректность ответа

Успешность интеграций

Задержки и ход диалога

Task Success Rate показывает долю задач, доведенных до результата

Локальные метрики показывают, где и почему возникают потери



Показатели пилотных проектов

Автоматизация

50-75%

Доля диалогов без участия сотрудника*

Task Success Rate

65-85%

Доля задач, доведенных до согласованного результата*
Включая предусмотренную маршрутизацию

Анализ **Task Success Rate** позволил точнее оценить качество обслуживания и найти причины нерешённых задач



* Автоматизация оценивается по диалогам, Task Success Rate — по задачам. Разность процентов не равна доле нерешенных задач.

Спасибо за внимание!

Москва

105120, 2-ой Сыромятнический пер., 1
+7 495 669 7440
stc-msk@speechpro.com

Санкт-Петербург

195277, Выборгская наб., 45, лит. Е
+7 812 325 8848
stc-spb@speechpro.com