

Больше, чем FAQ:

как RAG снизил нагрузку на операторов ОТП Банка

Антон Гусев, владелец продукта AI Роботы

Денис Загороднев, ML - аналитик

ОТП Банк: будущее начинается сейчас



Входит в международную OTP GROUP
30+ лет работает в России
ТОП банков по масштабу

Клиенты

- Реальные потребности
- Внимание и уважение
- Быстро и честно

Продукты

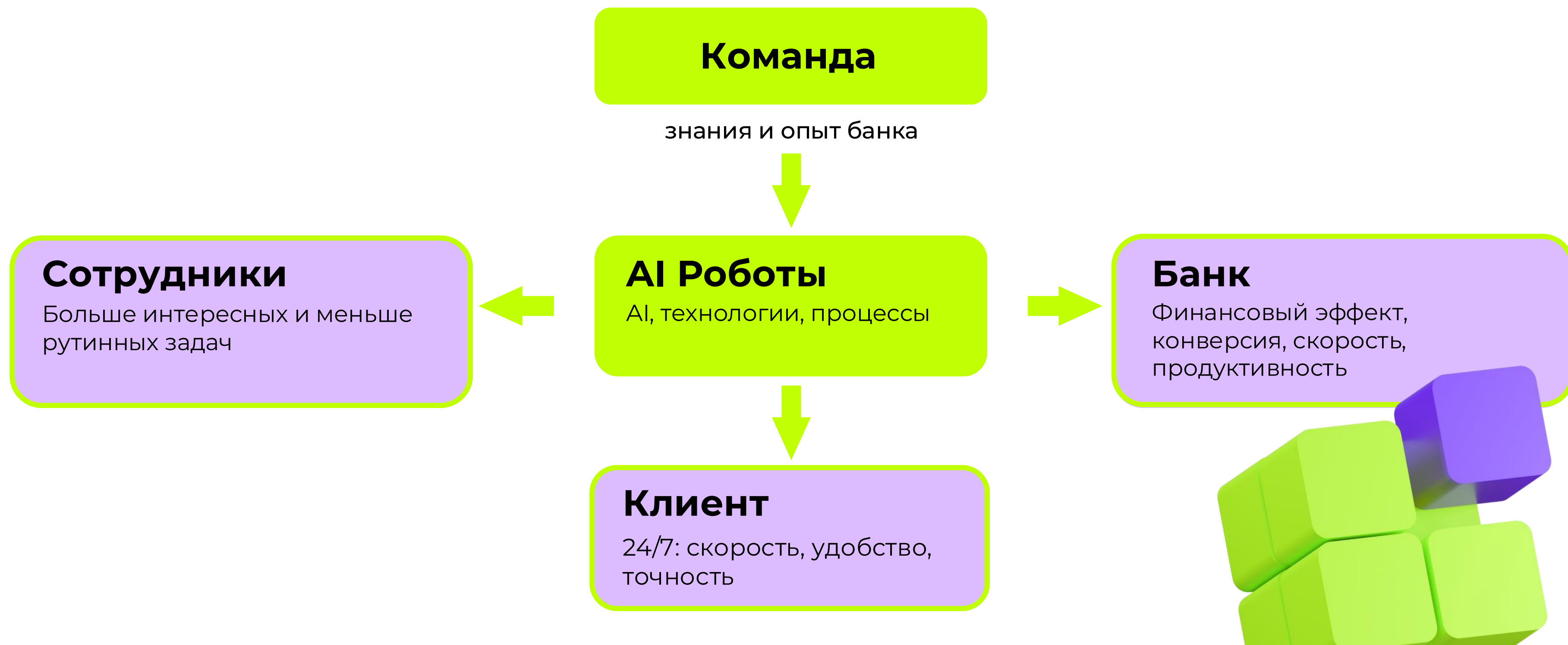
- Дистанционно и онлайн
- Точно и удачно
- Интернет и приложения

Инновации

- Автоматизация
- AI, LLM, NLP
- Большие данные

Знания Банка для любого клиента

Сегодня банк конкурирует не только продуктами — он конкурирует качеством цифрового клиентского опыта



Пользователь думает не интентами

- "Как выбрать кэшбэк?"

- "Как подключить бонусы!"

- "Где выбирать Cashback в мобильном приложении?"

- «Как предлагаются кеш бэки на месяц?»

Один смысловой кластер

Выбор категорий Cashback

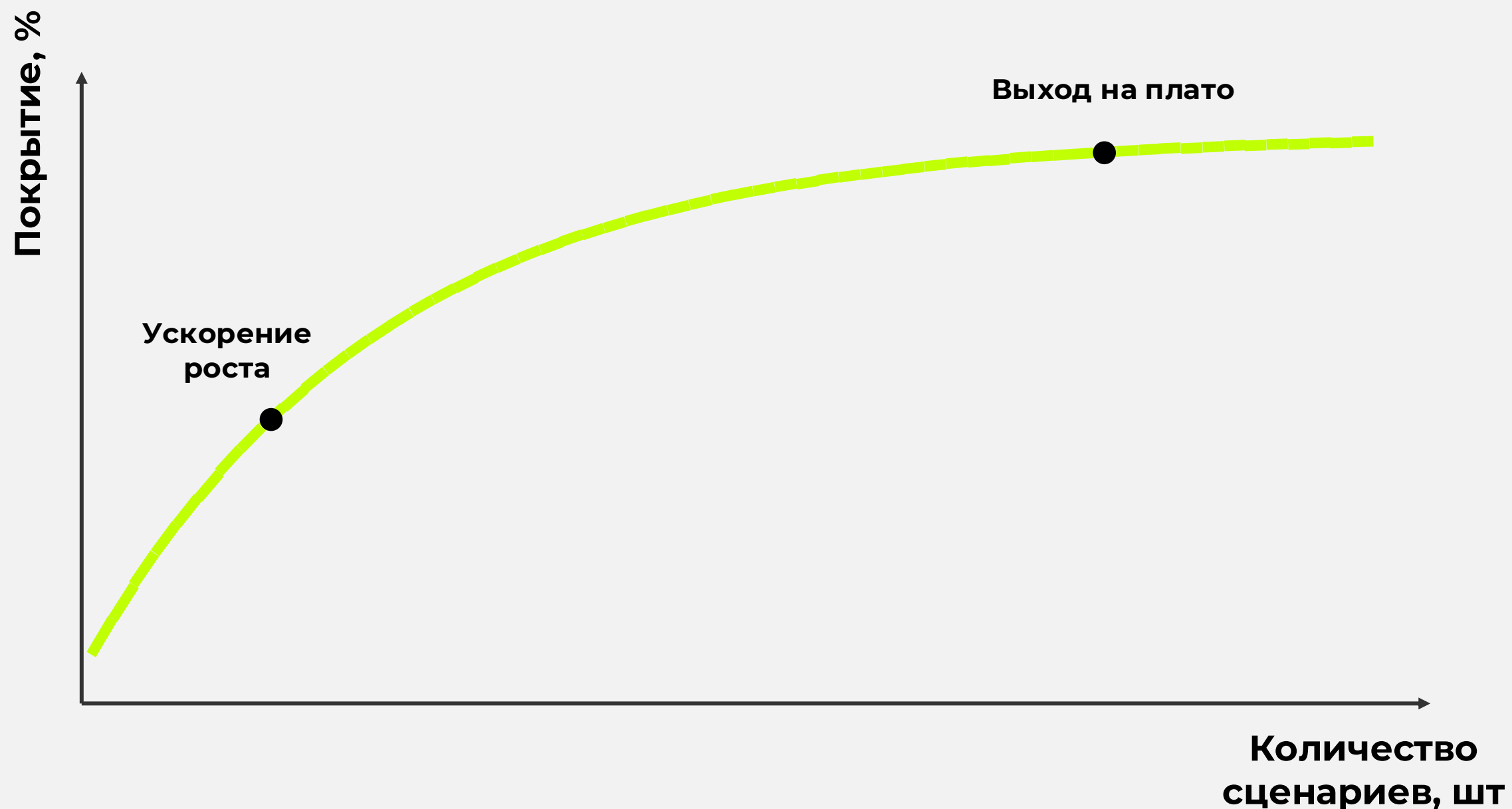
Условия

Процедура

Ограничения

Формулировки различаются. Нужное знание будет общим

Рост покрытия замедляется, поддержка усложняется

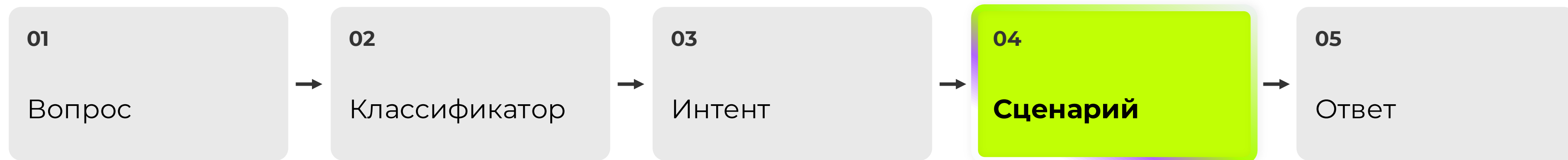


Сценарий требует:

- 1 Исследование**
Обращения, интенеты, условия
- 2 Разработка**
Ответы, тестирование, релиз
- 3 Поддержка**
Пересечения и устаревание

От сценариев к знаниям

1 Раньше: добавляем сценарий



2 Сейчас: добавляем знание



Теперь вопрос прямо связывается с информацией, нужной для ответа!

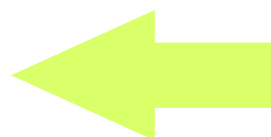
RAG или дообучение?

LLM

Работа с запросом и ответом

- Понимает запрос
- Использует контекст
- Формирует ответ

контекст



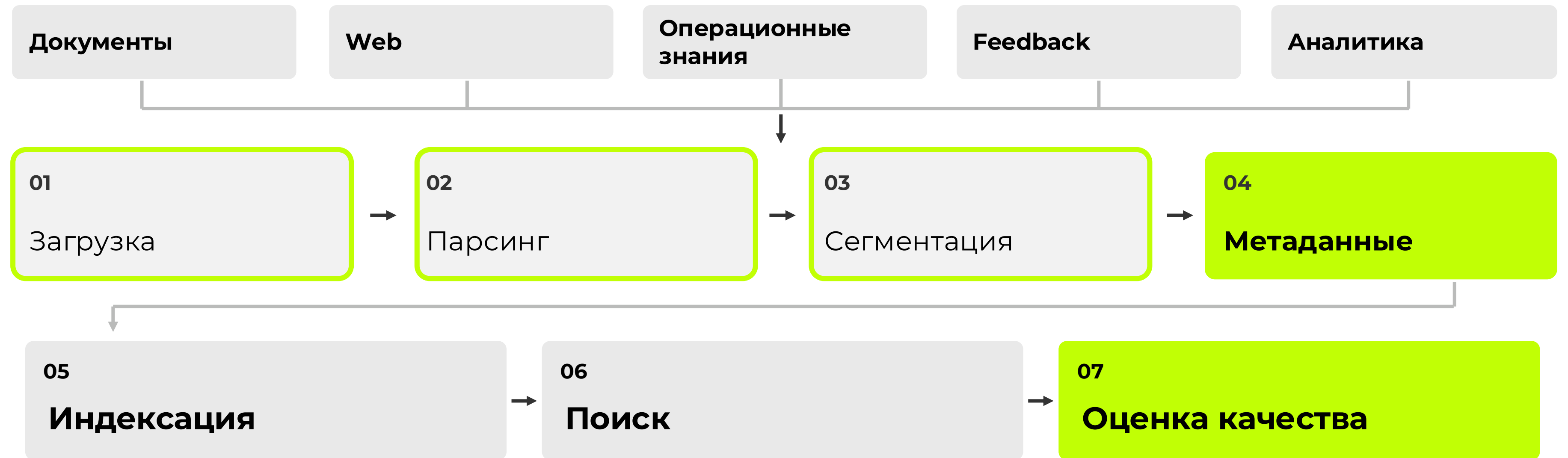
RAG

- 1 Хранит доступное знание
- 2 Находит нужную информацию
- 3 Позволяет отследить источник



Изменилось правило → обновляем версию знания

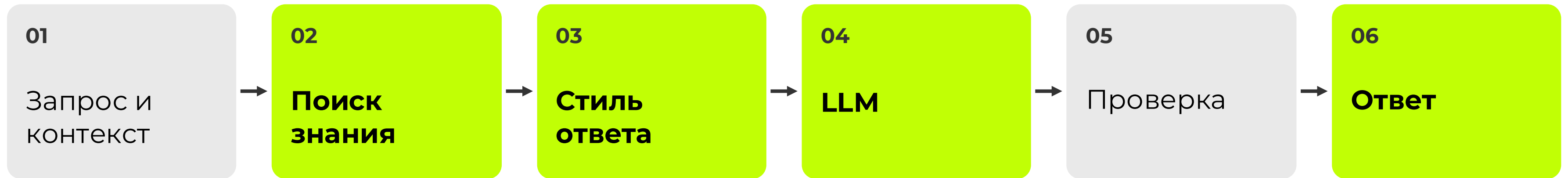
Хороший ответ – не просто документ...



Оценим качество найденного контекста:

- Источник
- Версия
- Актуальность
- Применимость

Хороший ответ – знание, стиль и проверка!

**1**

Что сказать

Knowledge / RAG

- Релевантные знания
- Контекст ситуации

2

Как сказать

Tone of Voice

- История диалогов
- Тон и детализация

3

Как сформулировать

LLM

- Естественный ответ
- Опора на контекст

Tone of Voice: стандарты общения с клиентом

Каждый ответ LLM по правилам «Tone of Voice» - понятно, вежливо и последовательно с учетом контекста, независимо от ситуации:

- Язык - орфография, пунктуация, структура с сохранением смысла.
- Стилль - отсутствие канцеляризов, единый тон банка.
- Персонализация - корректное обращение к клиенту по имени.
- Структура - инструкции, Hold, безопасность, маршрутизация и другие типовые блоки.
- Последовательность - одинаковые принципы коммуникации во всех чатах.

Выгода:

Банк

- Единый опыт
- Стандартно и масштабно
- Ниже риск и выше лояльность

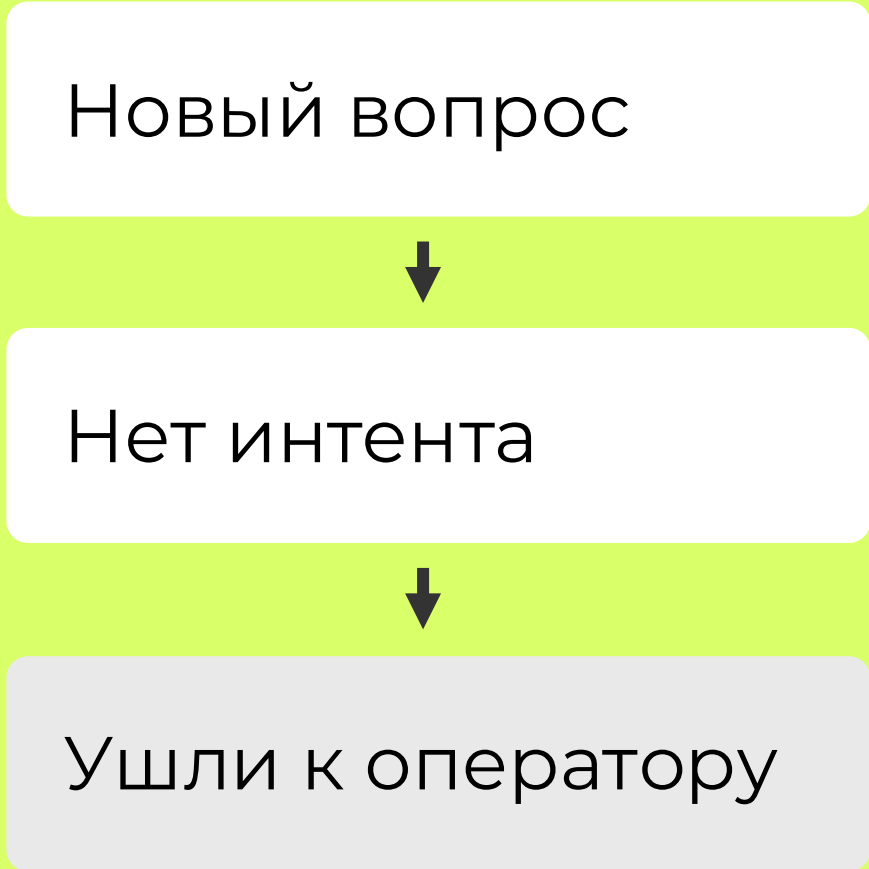


Клиенты

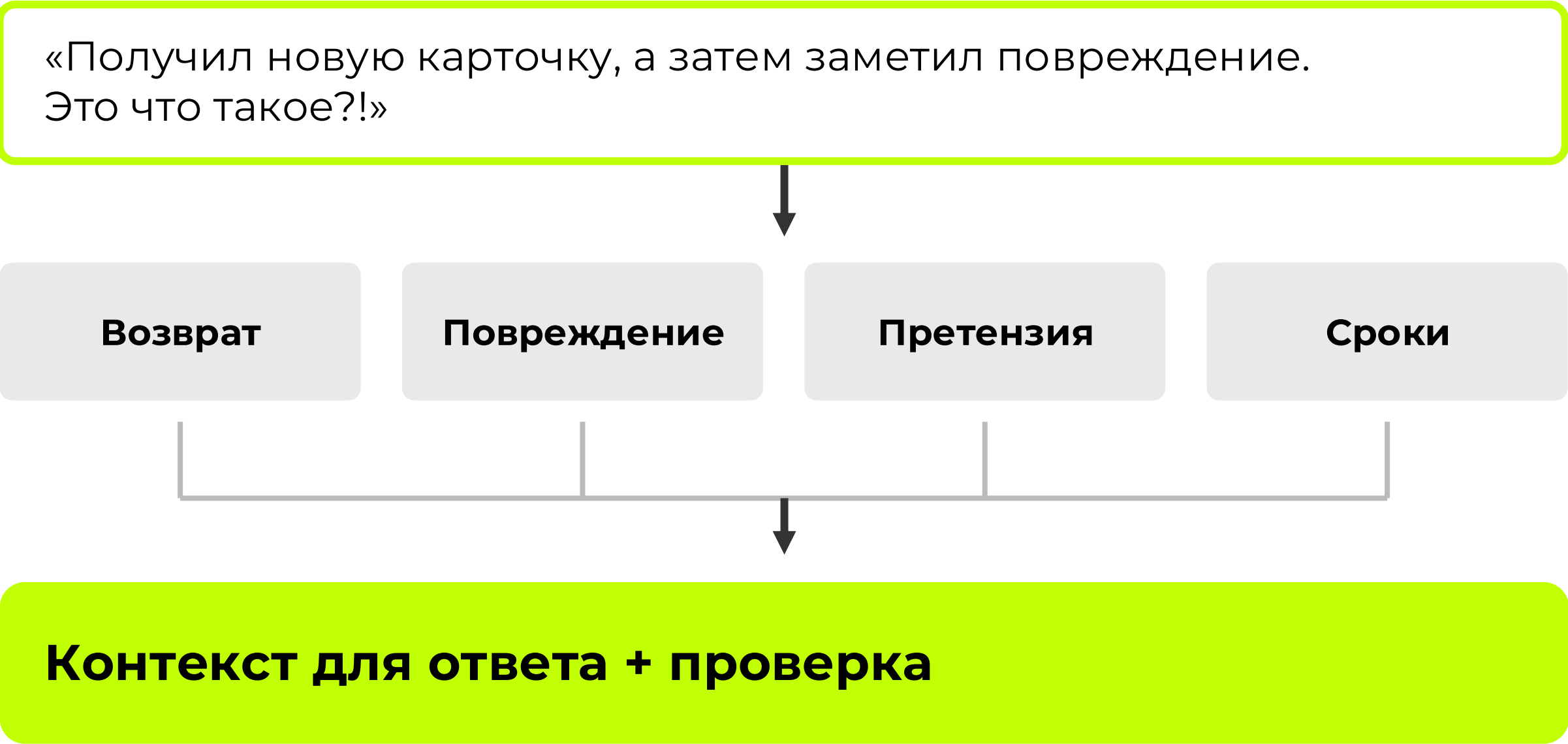
- Понятно и прозрачно
- Персонально и уважительно
- Доверительно

«Нет интента» – теперь не «нет ответа»

Раньше



Сейчас: ответим на вопрос, для которого в базе нет сценария



Качество проверяем на трёх уровнях

01

Поиск

Найдено нужное знание?

- Релевантность
- Актуальность
- Время поиска



02

Ответ

Знание использовано верно?

- Корректность
- Полнота
- Опора на источники



03

Результат

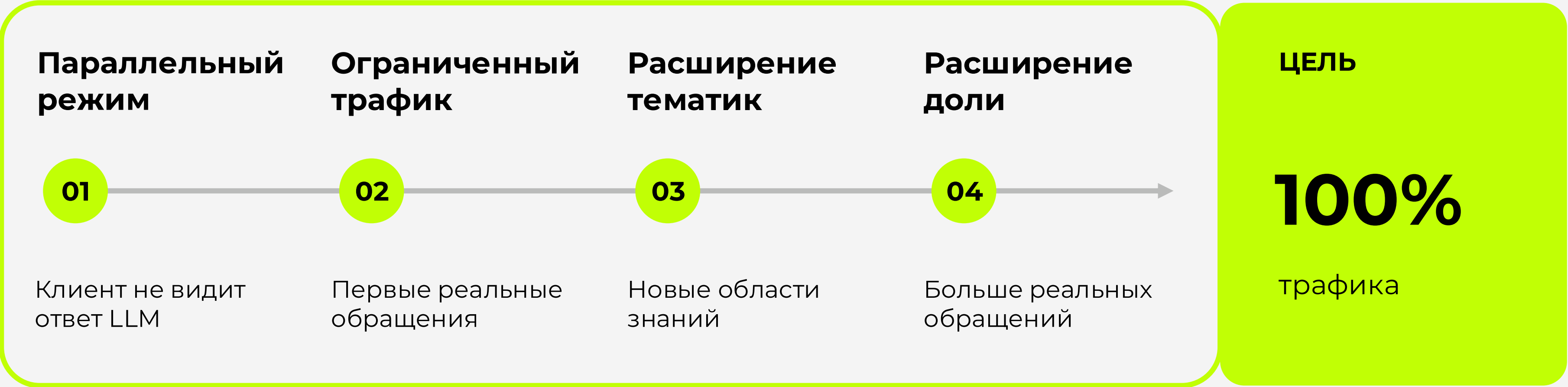
Задача пользователя решена?

- CSI
- Повторные обращения
- Автоматизация

Место ошибки определяет способ исправления

Уровень		Что произошло		Что улучшать
01	Поиск	Нужное знание не найдено	→	Индексация и поиск
02	Поиск	Найдена старая версия	→	Версии и актуальность
03	Поиск	Источники противоречат	→	Приоритет источников
04	Генерация	Ответ не подтверждён знанием	→	Проверка по источникам
05	Генерация	Ответ не подходит ситуации	→	Контекст и стиль

К 100% трафика идём через этапы проверки



Нам поможет:

- Мониторинг
- Валидация
- Маршрутизация
- Резервные сценарии

Эффект оцениваем по решённым обращениям

Эффект автоматизации

Направления оценки

Результаты требуют измерений

01

Пользователь

- Меньше тупиков
- Меньше повторных обращений

02

Бизнес

- Больше автоматизированных обращений
- Переиспользование знаний

03

Команда

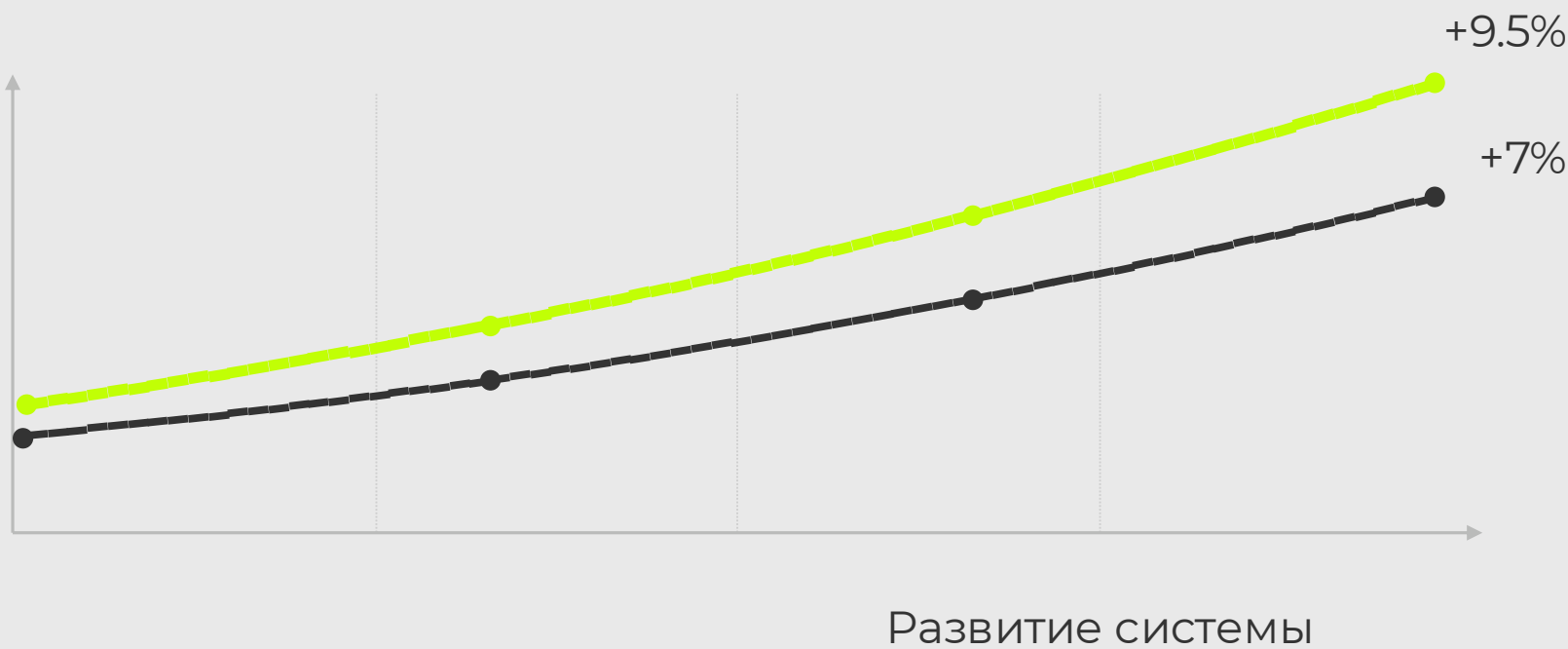
- Обновление знаний отдельно от модели
- Меньше ручного проектирования сценариев

Бизнес-эффект: показатели

Совместный рост модельных и бизнес-метрик

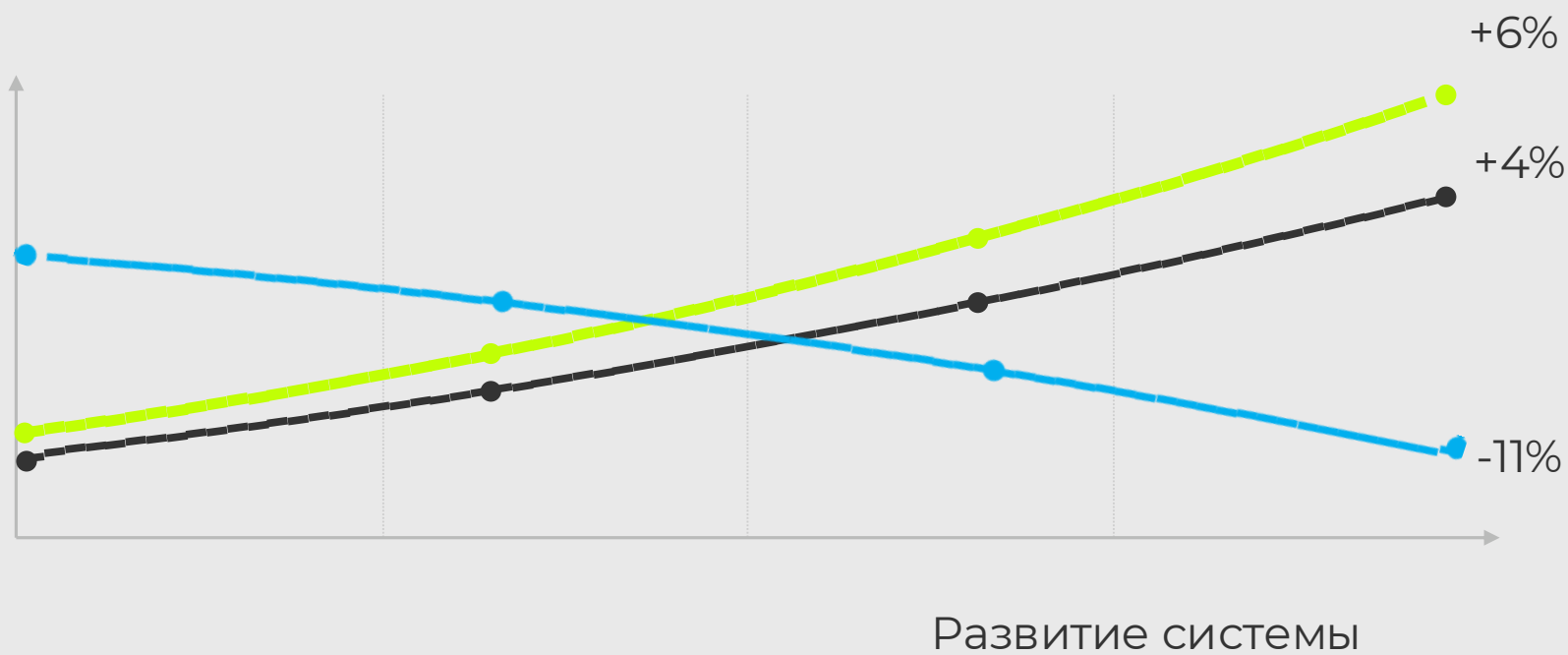
01 Модельные метрики

- Покрытие запросов
- Точность предсказания



02 Бизнес-метрики

- Автоматизация
- Уровень удовлетворенности
- Затраты на покрытие



Рост модельных метрик явно отражается в бизнес-показателях

Планы применения технологий

LLM и RAG – не только Чат-Бот. Это удачные и полезные процессы:



LLM превращает неструктурированный текст в понятную информацию, структурированные данные и автоматизированные действия!

От сценариев к знаниям

Новому вопросу нужно...

подходящее знание!

01

Найти



02

Применить



03

Проверить

Мы строим систему, которая умеет работать со знаниями



Антон Гусев,

владелец продукта AI Роботы

A.Gusev@otpbank.ru

Денис Загороднев,

ML – аналитик

D.Zagorodnev@otpbank.ru

