

Использование LLM в задачах:

путь от хайпа

к реальным решениям

Евгений Щербаков

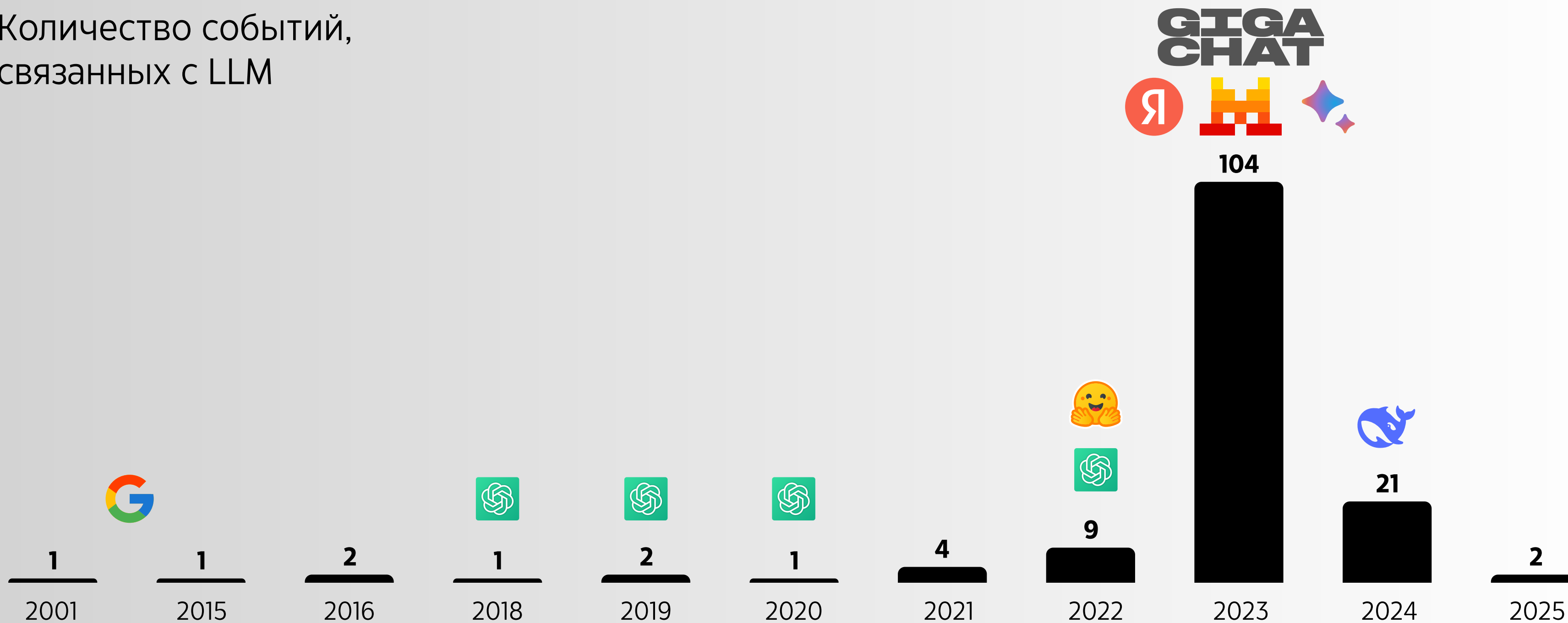
Бизнес партнёр по IT
Digital Operator | Ростелеком КЦ



Почему LLM — это не просто хайп?



Количество событий,
связанных с LLM



От идеи к реализации:

как мы начинали



На дворе был 2023 год ...



Ничего не понятно, но очень интересно



Волшебная пилюля, которая заменит кучу людей



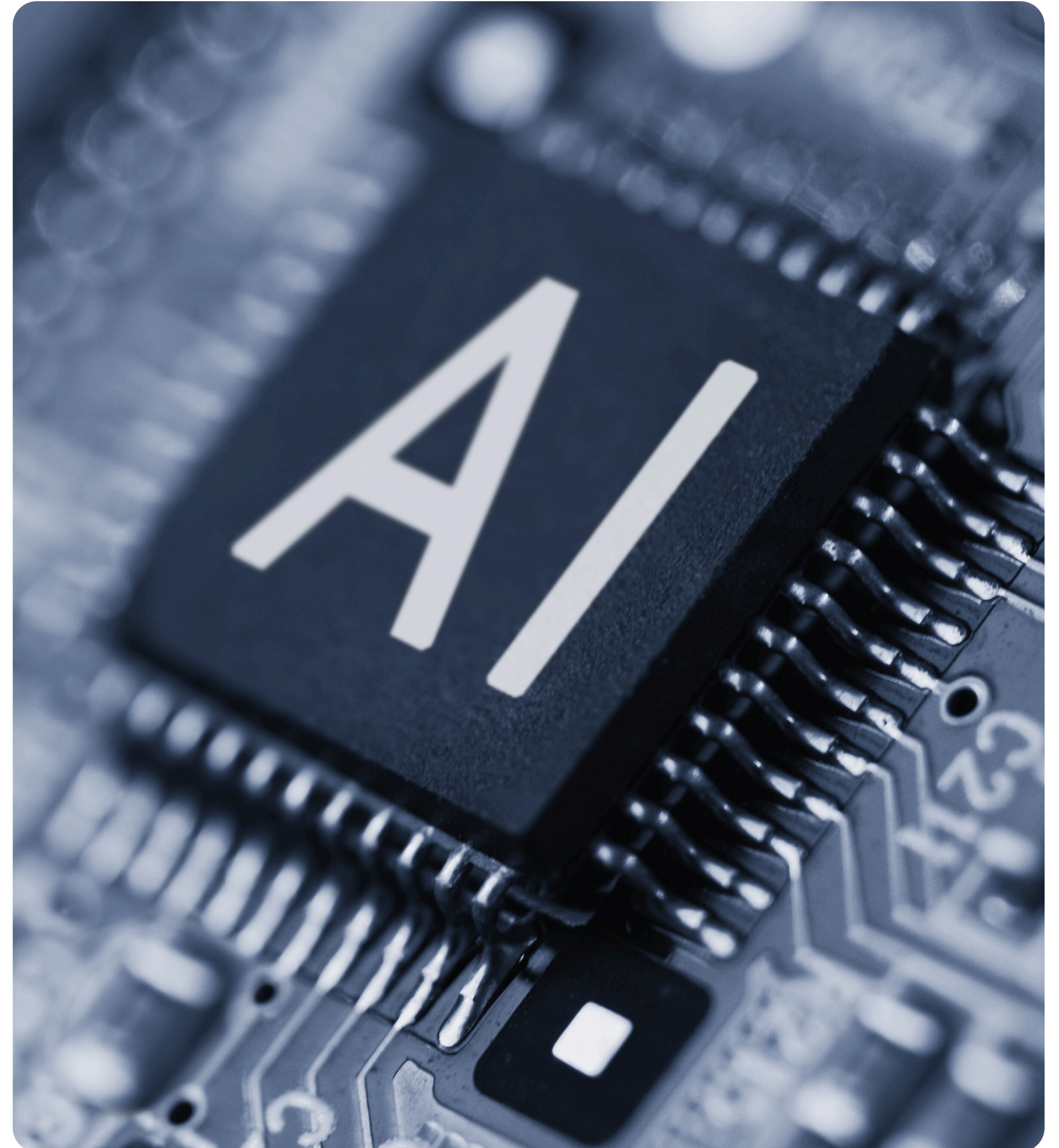
Ограничения к использованию мировых технологий



Амбициозный вызов:

какие задачи мы хотим решить

- ➔ Ассистент для операторов на основе базы знаний
- ➔ Не алгоритмичные роботы на свободных вопросах в тексте и голосе
- ➔ Нормализация текста и улучшение комментариев
- ➔ Классификация тематик

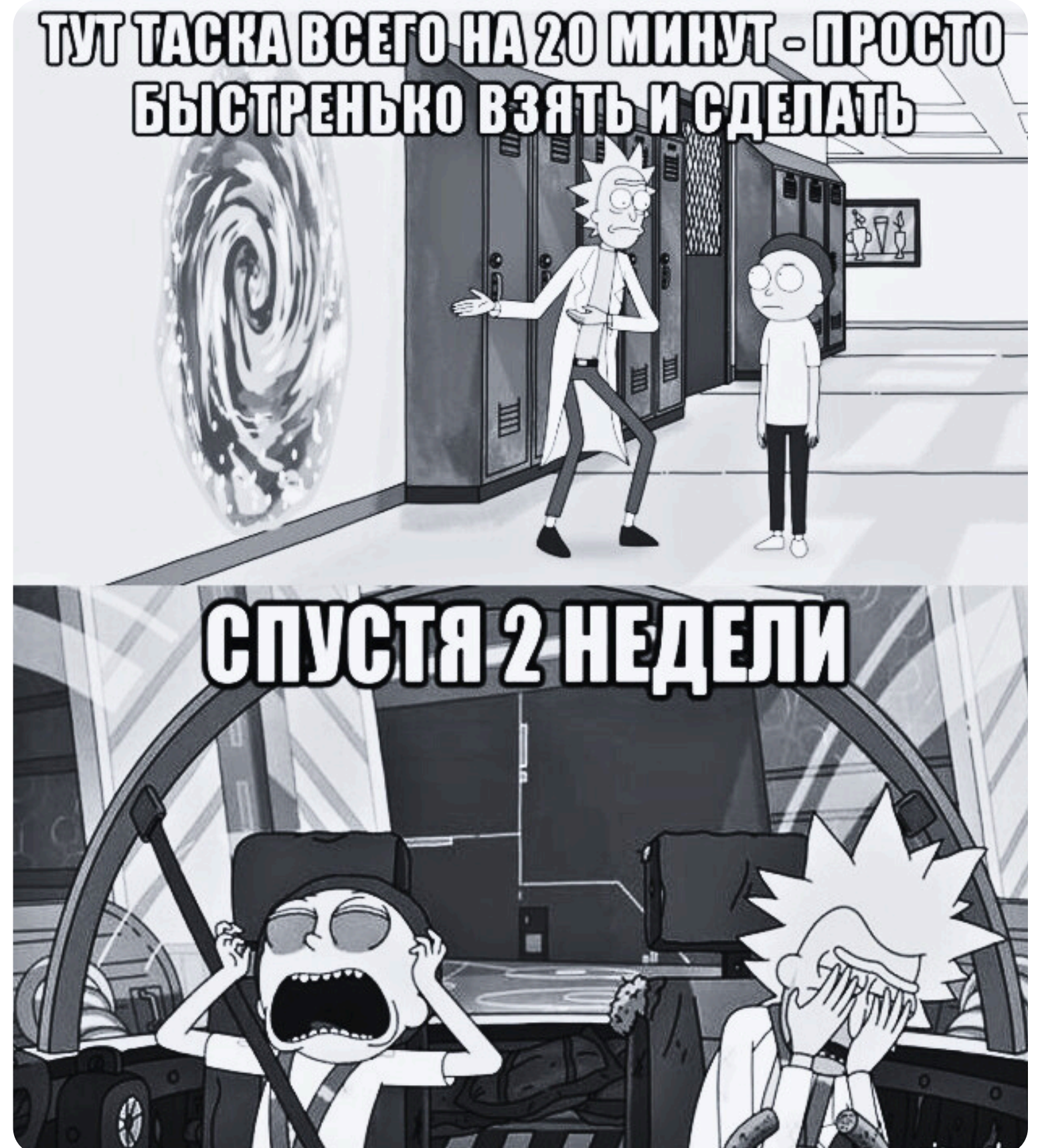




Решили попробовать...



Работы займут немного времени



Что имеем «на руках»



Внешние сервисы за пределами РФ

- ✓ Лучшая технология
- ✗ Зависимость от внешних сервисов
- ✗ Ограничение доступа к технологии
- ✗ Ограничения для работы с чувствительными данными

Сервисы в RU-сегменте

- ✓ Готовые сервисы приемлемого качества
- ✗ Зависимость от внешних сервисов
- ✗ Ограничения для работы с чувствительными данными

On-Prem установка

- ✓ Контролируемая среда сервиса
- ✓ Безопасность решения
- ✗ Высокие требования к внутренней экспертизе
- ✗ Высокие требования к инфраструктуре

Как погрузить

МОДЕЛЬ В КОНТЕКСТ



Не могу войти в аккаунт.
Что мне делать?



Базовая модель



Если у вас возникли проблемы со входом, попробуйте следующие шаги:

1. Проверьте правильность данных
2. Очистите кеш браузера | приложения
- 3....

Не могу войти в аккаунт.
Что мне делать?



Дообученная модель



Сожалею, что возникли проблемы со входом. Вы можете попробовать сбросить пароль, используя опцию «Не помню пароль» на главной странице или создать обращение в поддержку.



ПРОМТ ИНЖИНИРИНГ

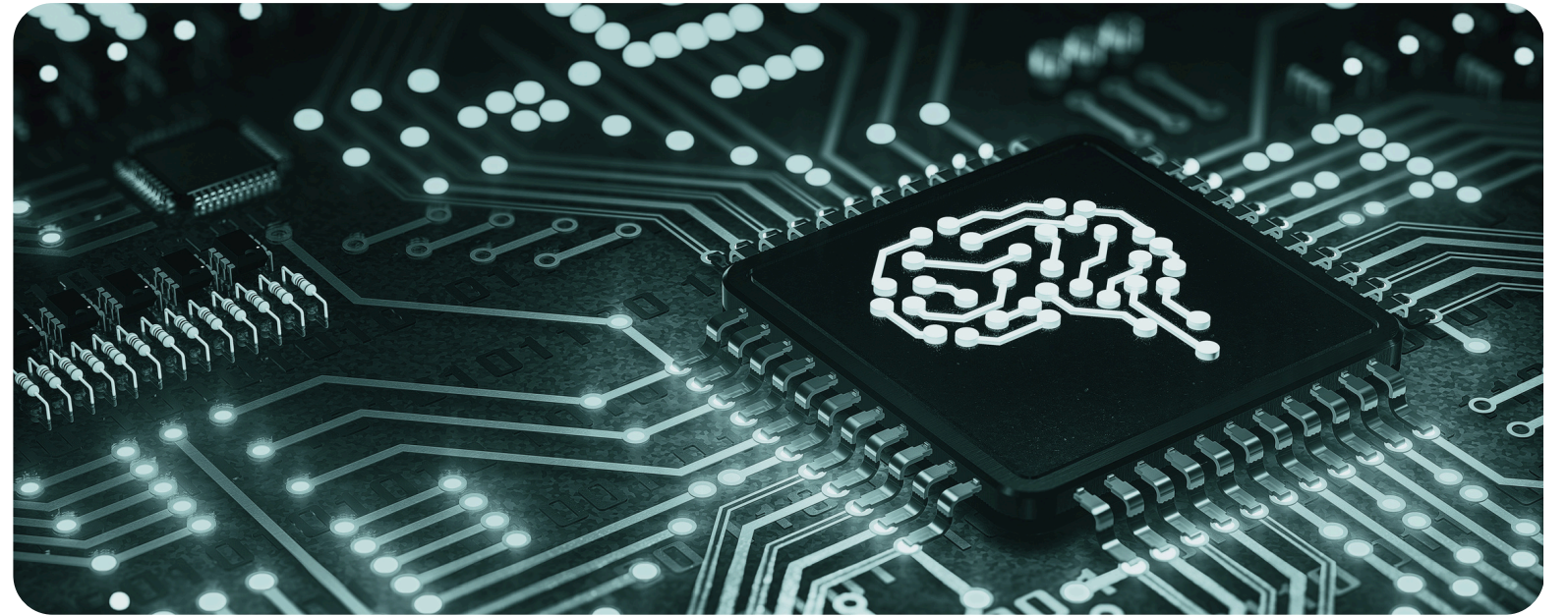
ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста



- 1** Выбрать модель и размер
- 2** Создать промт

Инструкция

Глоссарий

Примеры
Вопрос-ответ

- 3** Тестирование и аналитика
- 4** GOTO 1

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста

БЫЛО

ЮЛ создать на ГУ УЗ организации, не рук

СТАЛО

Обращается юридическое лицо. Хотели бы создать учетную запись организации на Госуслугах, но не является руководителем.

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста

БЫЛО

доб 0729 по погашение %

СТАЛО

Просят соединить со специалистом по добавочному 0729. Вопрос по погашению процентов.

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста

как сварить борщ?

Ассистент

Это не вопрос, связанный с деятельностью
понадобятся следующие ингредиенты:

или бизнесом. Это вопрос, связанный с кулинарией. Чтобы сварить бо

- * Свекла
- * Морковь
- * Лук
- * Картофель
- * Томаты
- * Капуста
- * Зелень
- * Сметана
- * Сахар
- * Соль
- * Перец

Приготовление борща включает в себя следующие этапы:

1. Очистите и нарежьте овощи.
2. Варите овощи в воде до тех пор, пока они не станут мягкими.
3. Добавьте томаты, капусту и зелень.
4. Варите еще несколько минут.
5. Добавьте сметану, сахар, соль и перец.
6. Варите еще несколько минут.
7. Подавайте горячим.

Но это не вопрос, который я могу помочь вам решить в рамках моих обязанностей как редактора текстов, связанных с
Если у вас есть вопросы, связанные с бизнесом или , я буду рад помочь.

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста



Точность

91%

Скорость ответа

2 сек

Удовлетворённость

80%

Ограничение контекста



RAG

Retrieval Augmented Generation

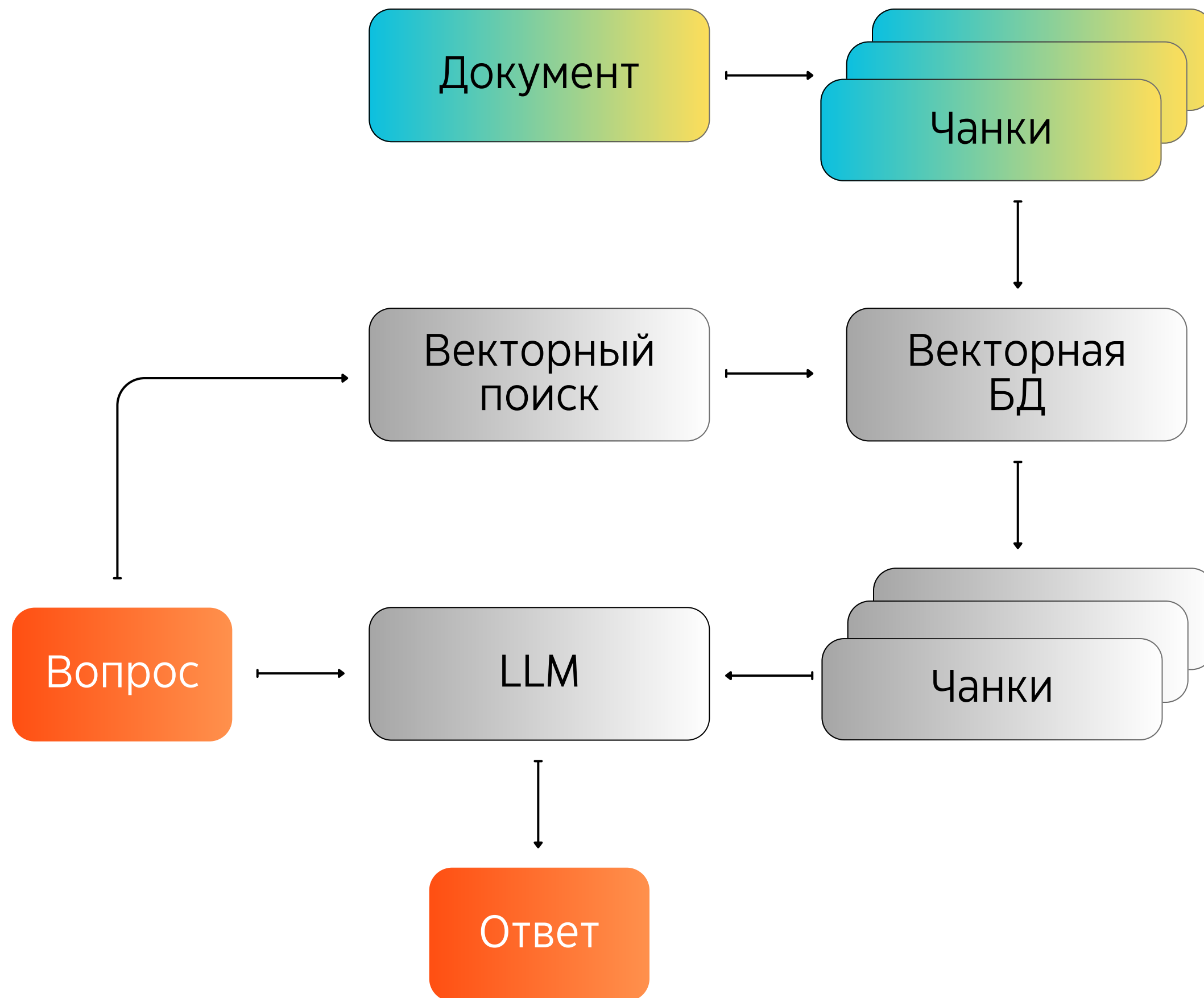
ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг
+ RAG

- ➔ Единообразии стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста
- ➔ Использование внешнего справочника



ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг
+ RAG

- ➔ Единообразии стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста
- ➔ Использование внешнего справочника

Перекрытие

RAG (Retrieval-Augmented Generation) — это метод, который комбинирует языковые модели с внешними источниками данных. Перед генерацией ответа модель ищет релевантные фрагменты в базе знаний, что позволяет ей использовать актуальную информацию. Этот подход снижает риск галлюцинаций и повышает точность ответов. Однако эффективность RAG зависит от качества чанкинга: размера чанков и перекрытий между ними. Мелкие чанки подходят для точного поиска фактов, а крупные — для сохранения контекста. Перекрытия между чанками помогают избежать потери смысла на стыках.

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг
+ RAG

- ➔ Единообразии стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста
- ➔ Использование внешнего справочника

#1

RAG (Retrieval-Augmented Generation) — это метод, который комбинирует языковые модели с внешними источниками данных. Перед генерацией ответа модель ищет релевантные фрагменты в базе знаний, что позволяет ей использовать актуальную информацию. Этот подход снижает риск галлюцинаций и повышает точность ответов.

#2

Этот подход снижает риск галлюцинаций и повышает точность ответов. Однако эффективность RAG зависит от качества чанкинга: размера чанков и перекрытий между ними. Мелкие чанки подходят для точного поиска фактов, а крупные — для сохранения контекста. Перекрытия между чанками помогают избежать потери смысла на стыках.

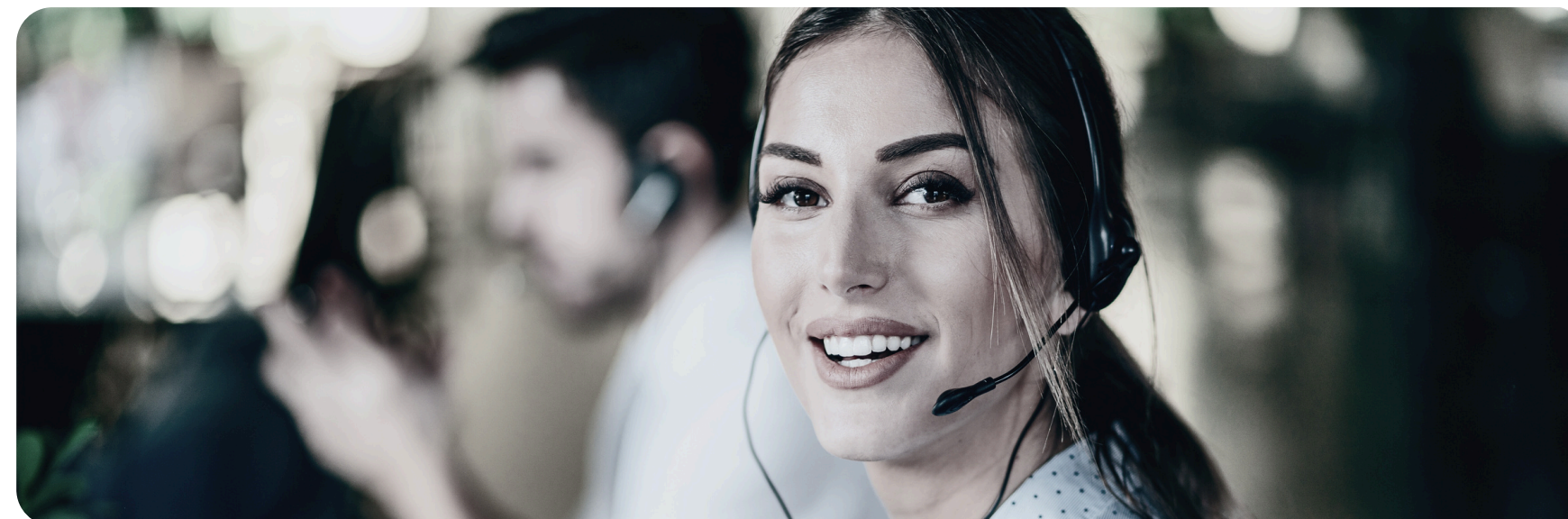
ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг
+ RAG

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста
- ➔ Использование внешнего справочника



БЫЛО

доб 0729 по погашение %

СТАЛО

Просят соединить с **Ивановым Иваном**, **кредитным специалистом**, по добавочному 0729. Вопрос по погашению процентов.

ПРИМЕР

улучшить качество

ответов операторов

Используем промт-инжиниринг
+ RAG

- ➔ Единообразие стилистики
- ➔ Расшифровка специфичных терминов
- ➔ Повышение детализации текста
- ➔ Использование внешнего справочника



Снижает зависимость от размера
контекстного окна

Отличный результат на справочниках

Сложный PipeLine



LoRA

Low-Rank Adaptation

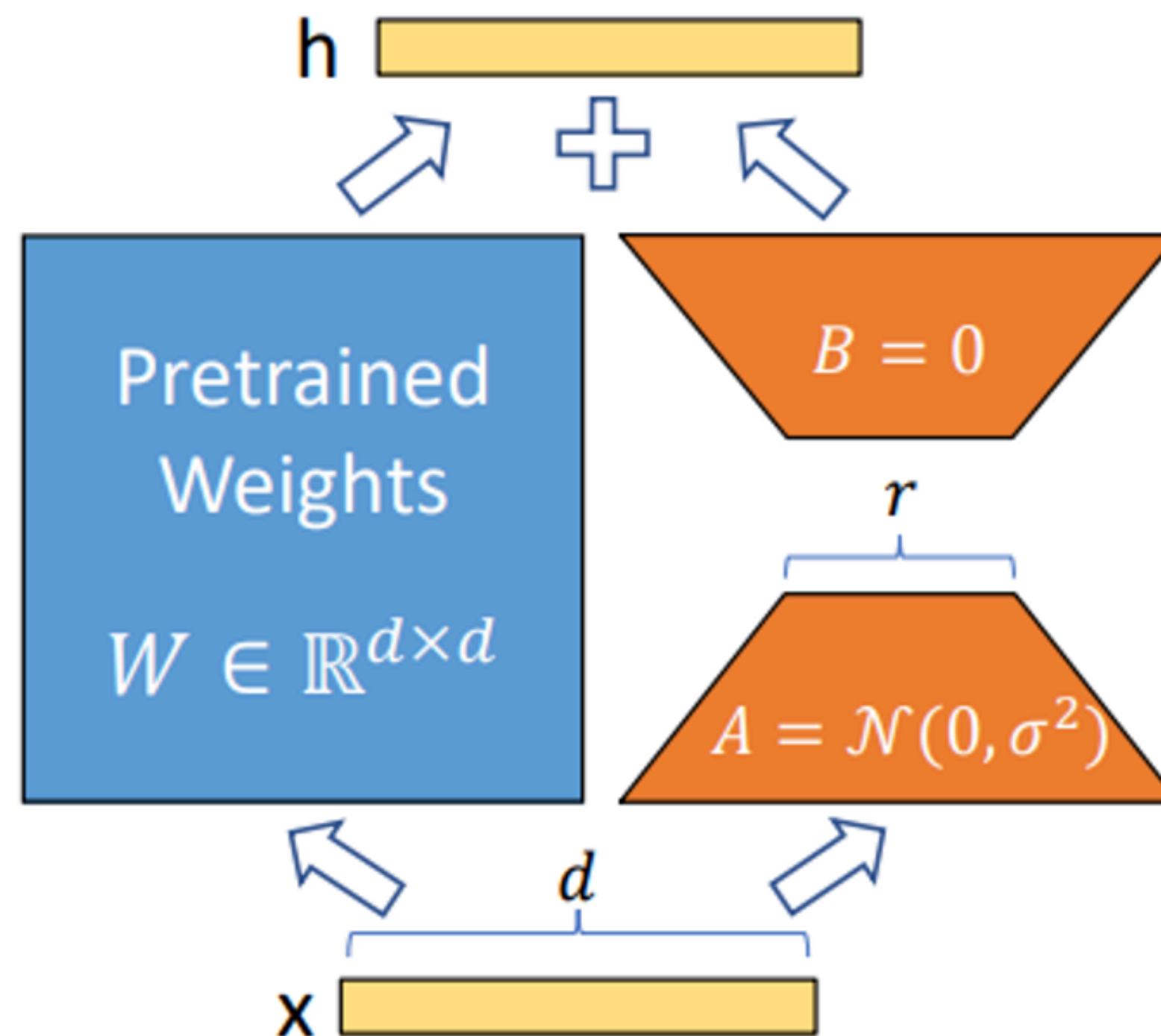
ПРИМЕР

улучшить качество

ассистентов

Используем промт-инжиниринг
+ RAG + LoRA

- ➔ Требуется меньше ресурсов по сравнению с полным обучением
- ➔ Формируется дополнительная матрица весов
- ➔ Основная матрица весов замораживается и не меняется



ПРИМЕР

улучшить качество

ассистентов

Используем промт-инжиниринг
+ RAG + LoRA

- ➔ Требуется меньше ресурсов по сравнению с полным обучением
- ➔ Формируется дополнительная матрица весов
- ➔ Основная матрица весов замораживается и не меняется

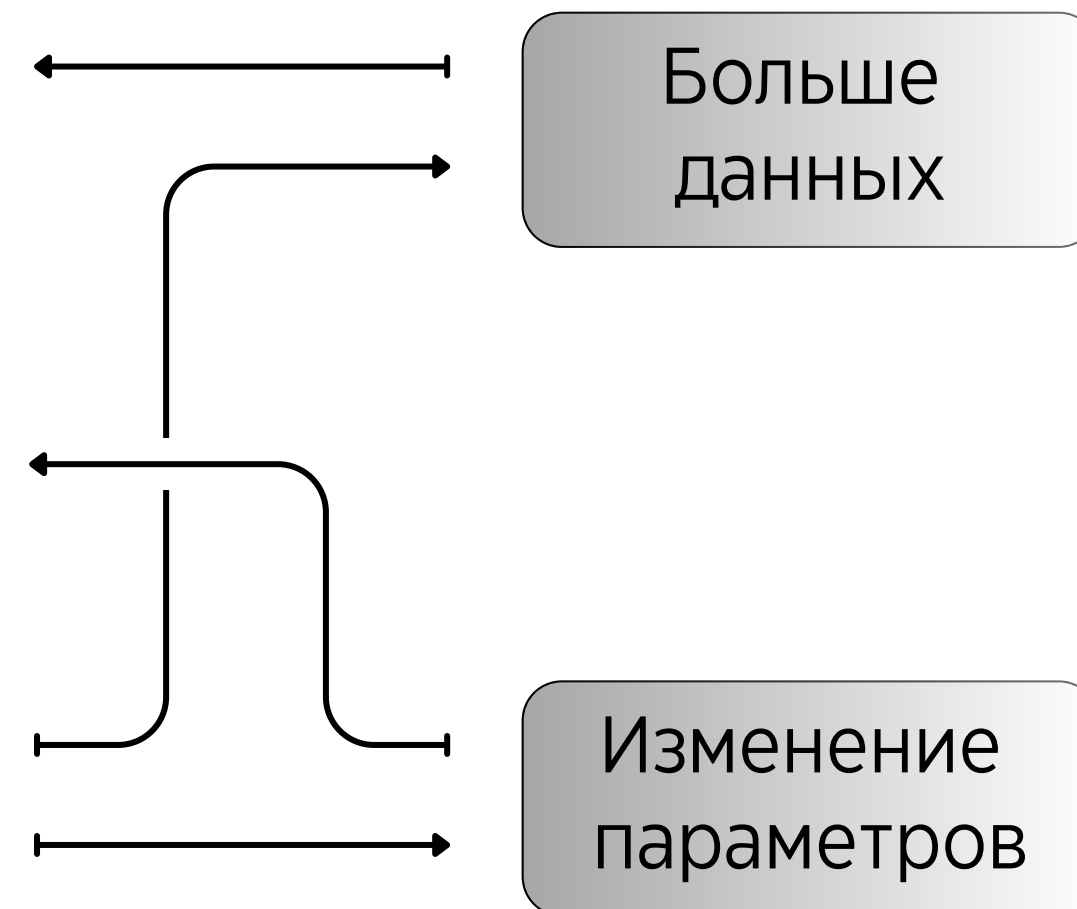
Подготовка данных
в формате JSONL

Создание сущности
датасета

Запуск
дообучения

Оценка
результатов

Интеграция



Расскажи пожалуйста про Зонтичный механизм предоставления поручительств. Я не понимаю как по нему консультировать.

Тест алгоритм

17.03.2025, 17:39:03

Зонтичный механизм предоставления поручительств - это упрощенный способ получить кредит, если у вас не хватает залога. **Банк** возьмет на себя основные риски перед банком, чтобы вы могли получить необходимое финансирование на развитие своего бизнеса.

Чтобы консультировать по этому механизму, вы можете сказать примерно следующее:

"Зонтичный механизм предоставления поручительств - это особый вид кредита, который позволяет получить финансирование без необходимости предоставления залога. Это особенно полезно для малого и среднего бизнеса, которые часто сталкиваются с проблемами при получении кредита из-за отсутствия достаточного залога.

При этом механизме **банк** выступает в качестве поручителя перед банком, что позволяет банку предоставить кредит без требования залога. Это означает, что вы можете получить необходимое финансирование для развития своего бизнеса, не имея при этом достаточного залога.

Условия получения кредита по зонтичному механизму следующие:

- * Поручительство покрывает 50% от суммы кредита;
- * Размер поручительства по кредиту (или нескольким кредитам) одного предпринимателя – до 1 млрд рублей;
- * Срок кредита не более 180 месяцев;
- * Цели кредита: инвестиционные цели, пополнение оборотных средств, развитие бизнеса, рефинансирование кредита.

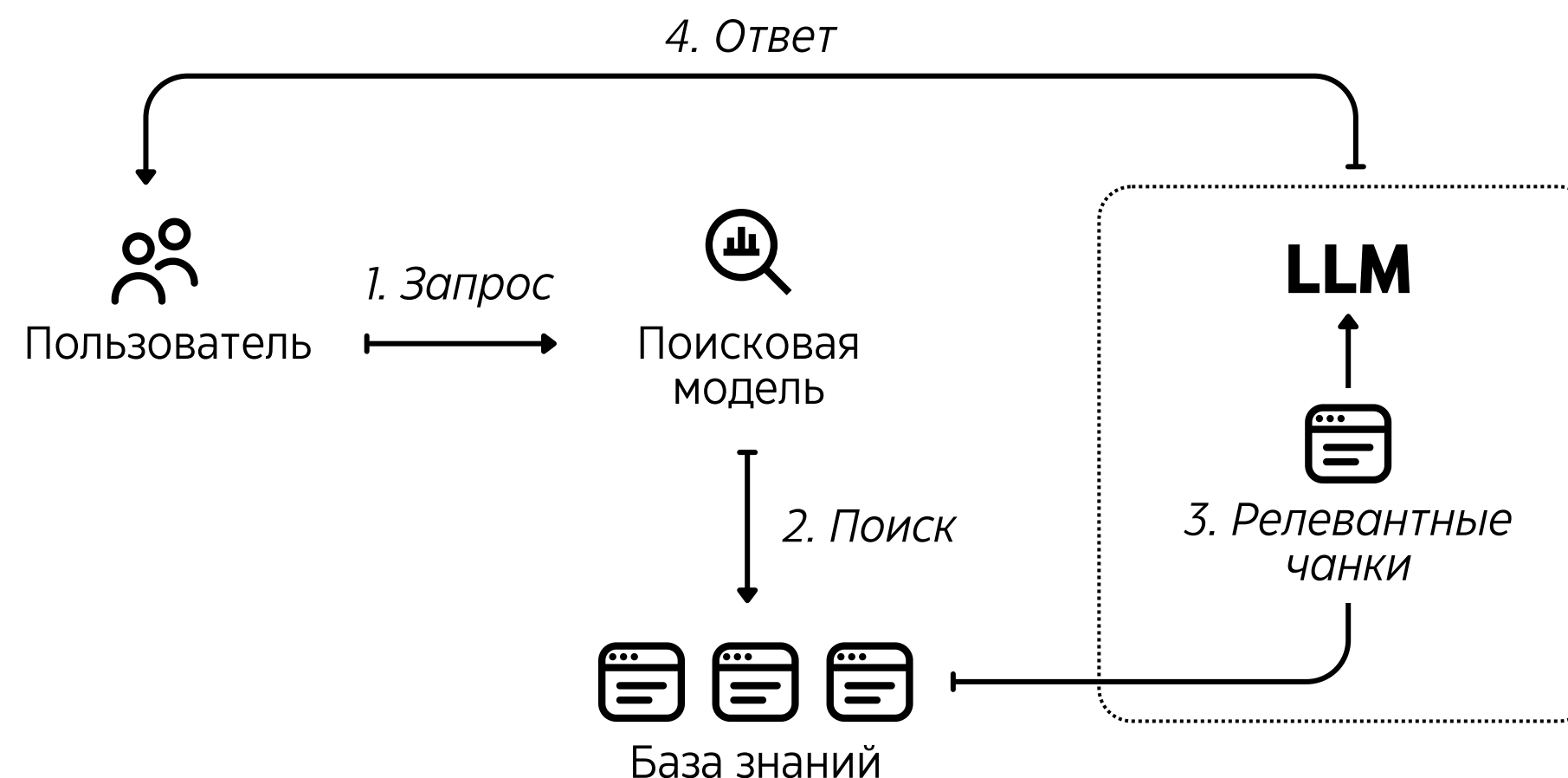
Если вы заинтересованы в получении кредита по зонтичному механизму, вы можете обратиться в банк, который сотрудничает с **банком**. Банк поможет вам подготовить все необходимые документы и подать заявку на кредит."

Таким образом, вы можете объяснить клиенту, что зонтичный механизм предоставления поручительств - это удобный и доступный способ получить кредит без необходимости предоставления залога, и что **банк** готова помочь

- ➔ LLM в чистом виде решает малое количество прикладных задач с нестабильным качеством.
- ➔ Для получения ощутимого бизнес-эффекта требуется комбинация систем и сервисов.
- ➔ При достижении успеха — не сложно тиражировать.

Рецепт успеха

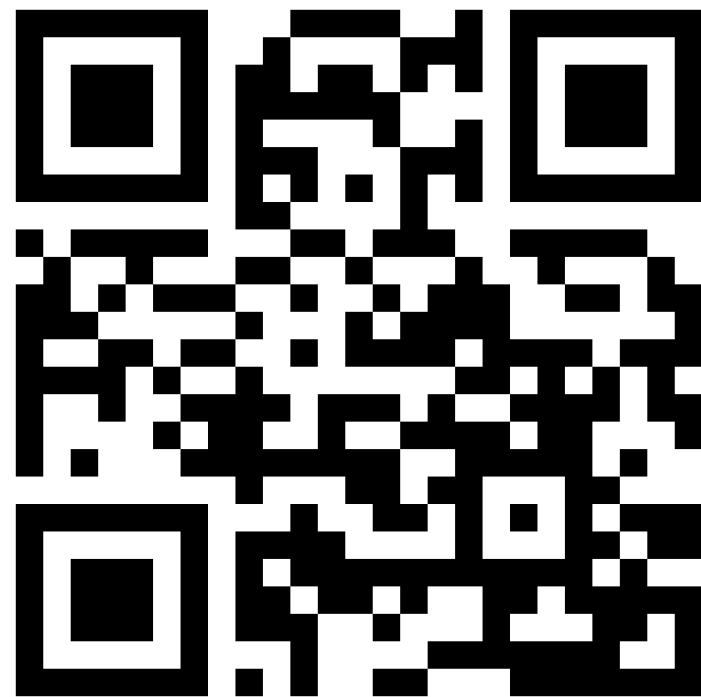
Вовлеченность
Инфраструктура
LLM + RAG + LoRA
Экспертиза




Спасибо

за внимание

Больше информации
на нашем сайте



 +7 499 550 0000

Контакты

 ccenter@rostelecom-cc.ru

