

КАК НАУЧИТЬ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ОБСЛУЖИВАТЬ КЛИЕНТОВ НЕ ХУЖЕ ЧЕЛОВЕКА?

Евгений Смирнов
Head of ML Laboratory & Chief Data Scientist
Альфа-Банк



О СЕБЕ

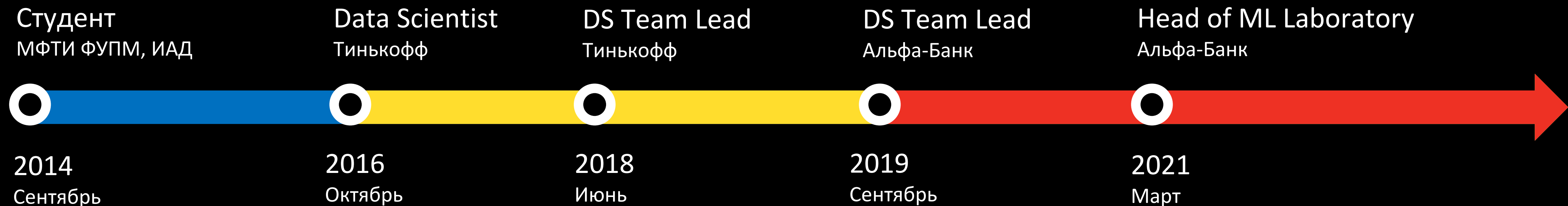


Евгений Смирнов

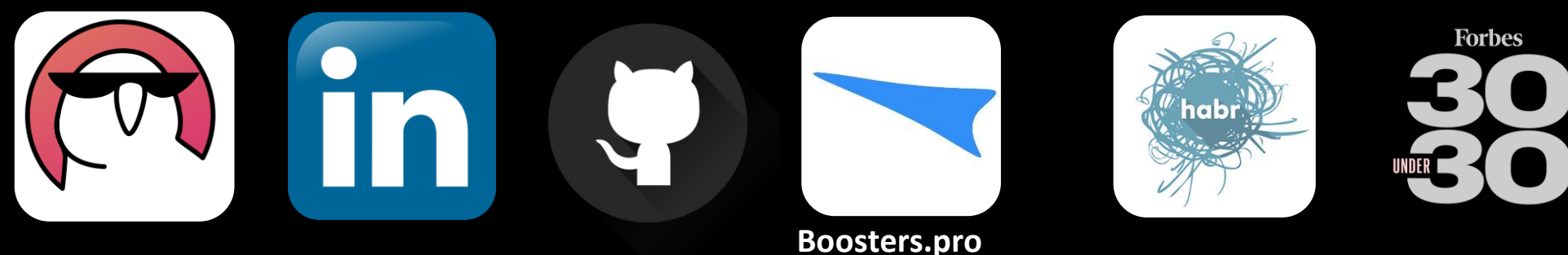
Head of Machine Learning Laboratory & Chief Data Scientist
ex Tinkoff, MIPT alumni

Автор Telegram-канала «Нескучный Data Science»
Хэдлинер курса «Принятие решений на основе данных»

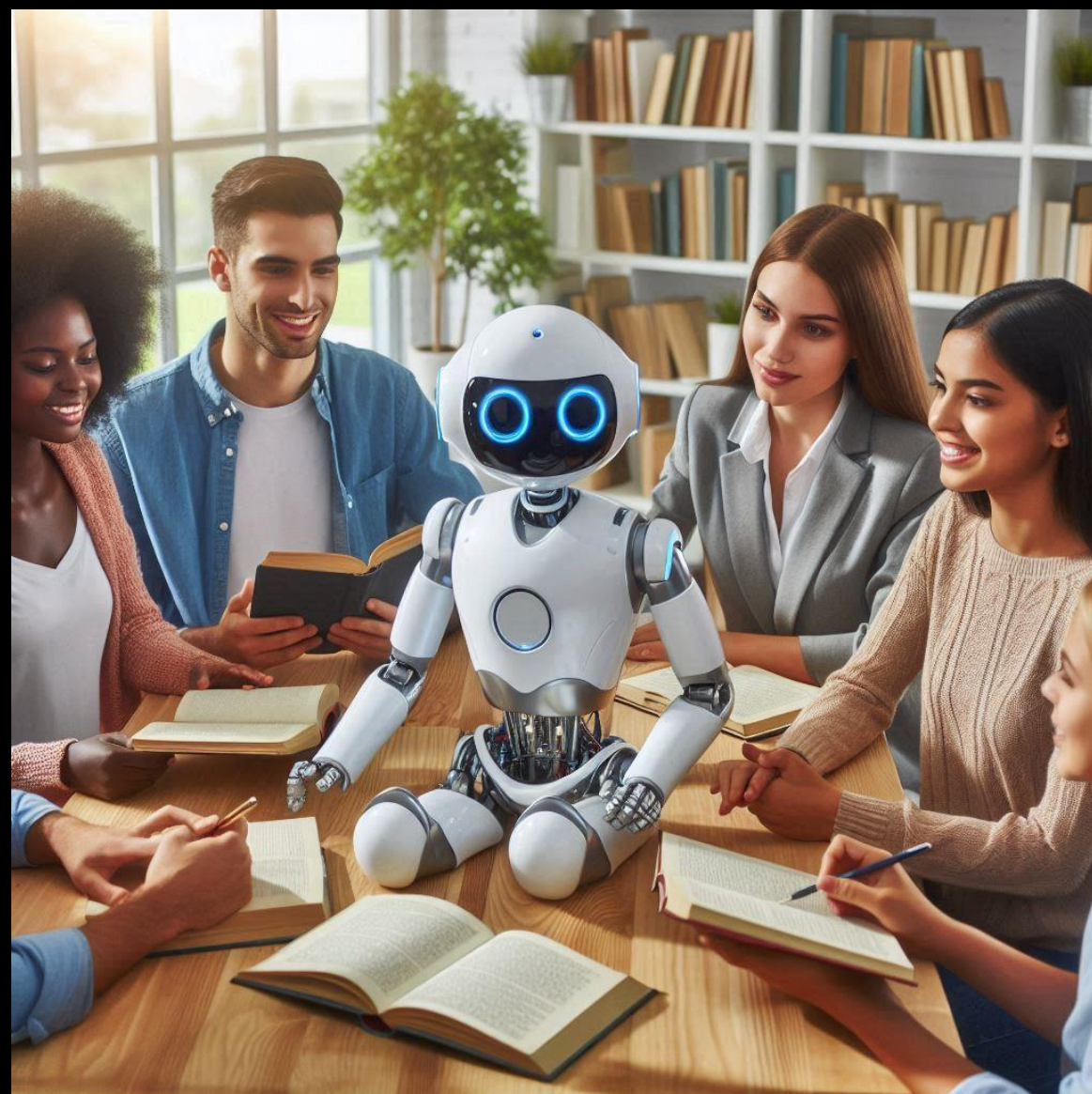
Карьерный путь:



Контакты и публикации:



ИИ ГОТОВ ВЗЯТЬ ОБЛУЖИВАНИЕ КЛИЕНТОВ НА СЕБЯ



ИИ и ЕИ обучаются и принимают решения по схожим принципам

+



обучения специалистов поддержки проходит быстро и среди них высокий уровень оттока

+



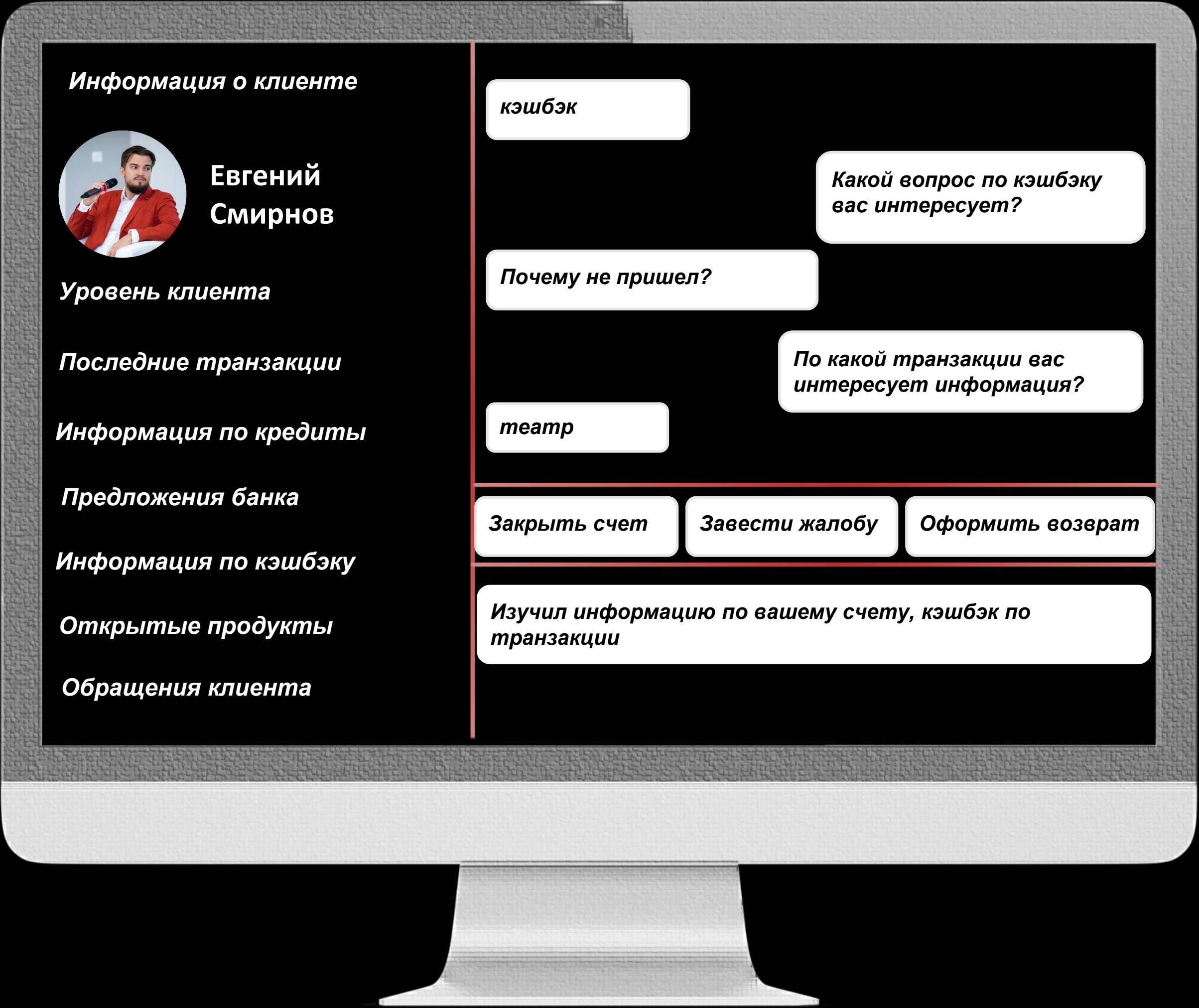
LLM решают задачи на уровне специалиста-середничка в разных областях

ПРОЦЕСС ОБСЛУЖИВАНИЯ КЛИЕНТОВ

1. Читает
вопрос клиента

2. Изучает
данные клиента

3. Вспоминает
информацию с
обучения



4. Задаёт
вопросы

5. Отвечает на
вопрос клиента

6. Совершает
целевые действия

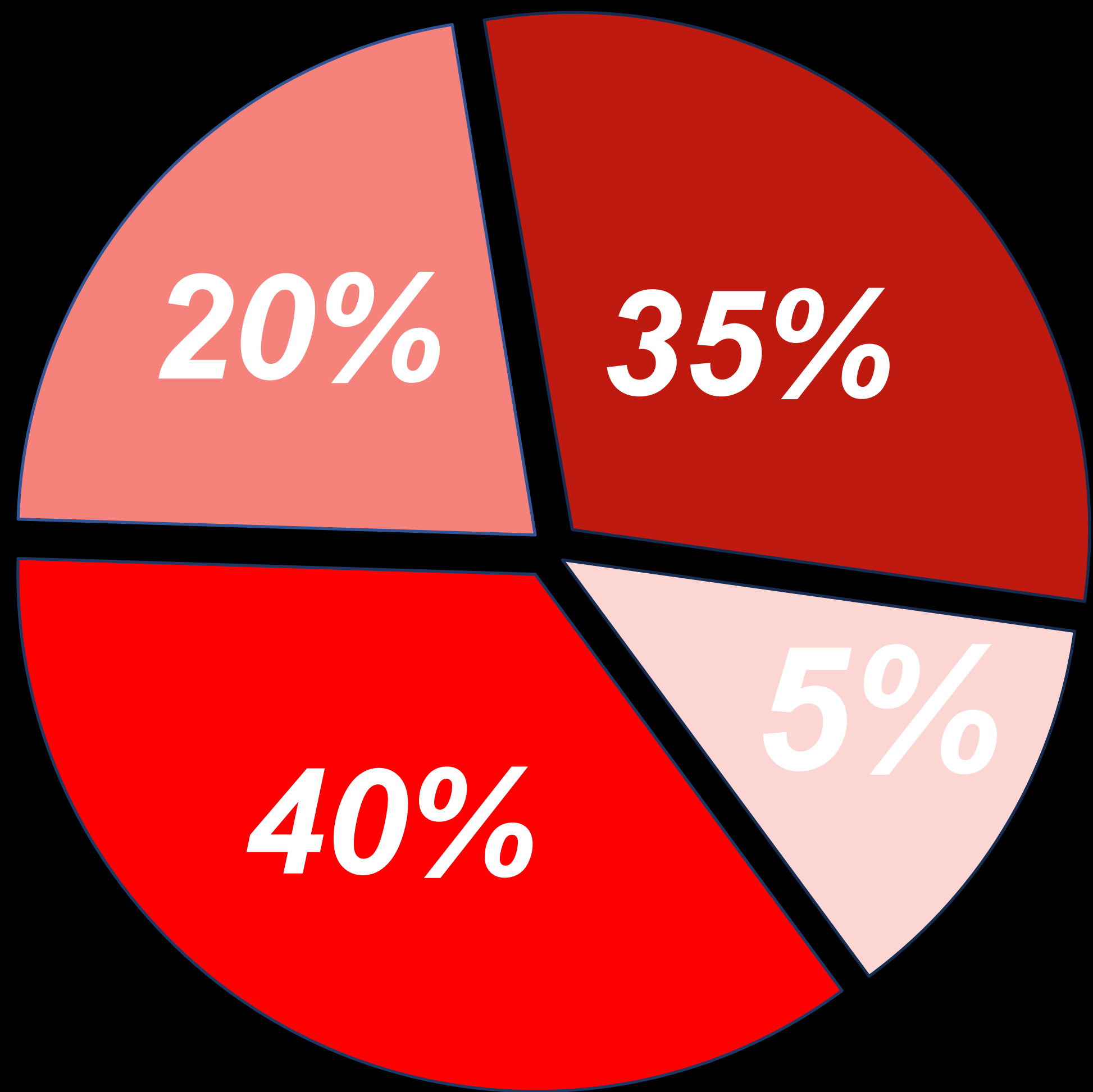
ВКЛАД КОМПОНЕНТ В АВТОМАТИЗАЦИЮ

**1. Изучает
данные клиента**

**Интеграция
данных**

**2. Вспоминает
информацию с
обучения**

RAG-пайплайн



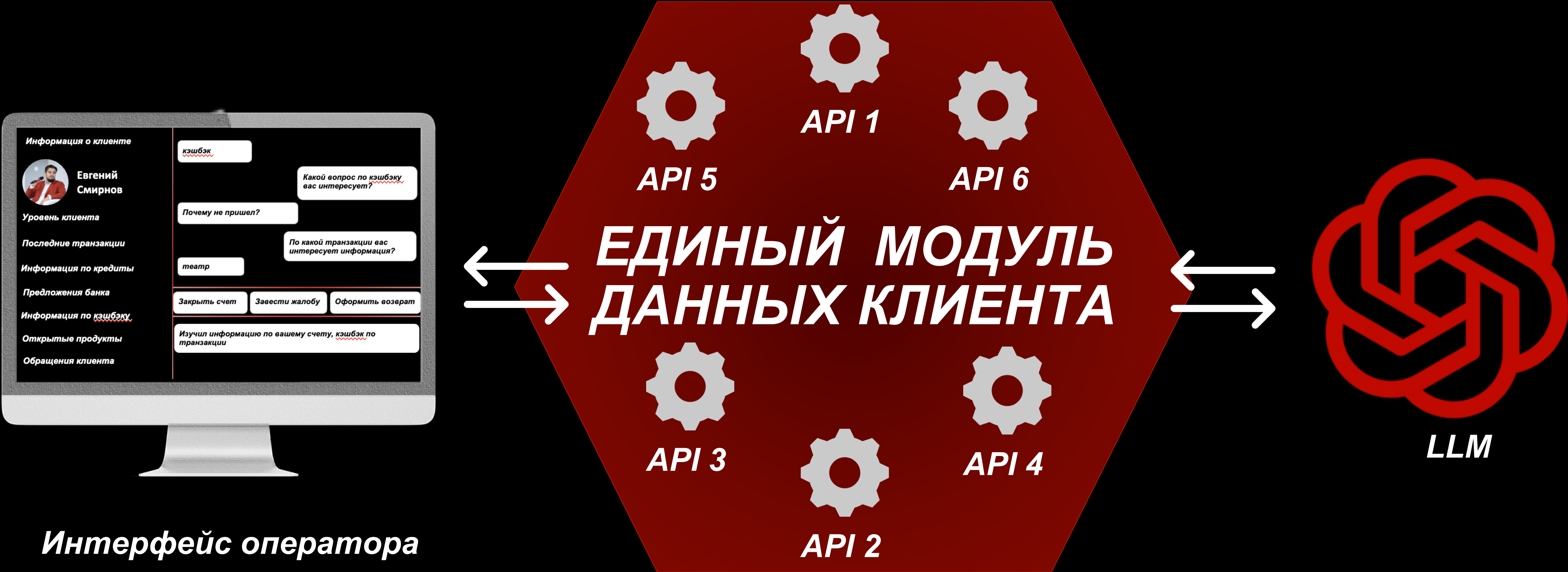
**3. Задает доп.
вопросы**

**Модуль
доспрашивания**

**4. Совершает
целевые действия**

**Классификатор
целевых действий**

ВКЛАД КОМПОНЕНТ В АВТОМАТИЗАЦИЮ



! LLM и оператору требуется одинаковое количество данных в одном и том же формате

ИСПОЛЬЗОВАНИЕ ДАННЫХ КЛИЕНТОВ В LLM

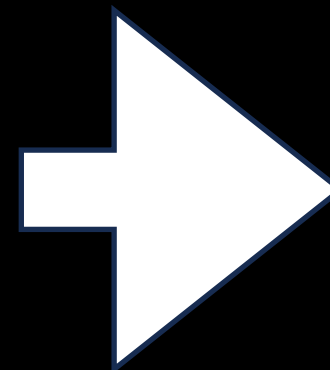
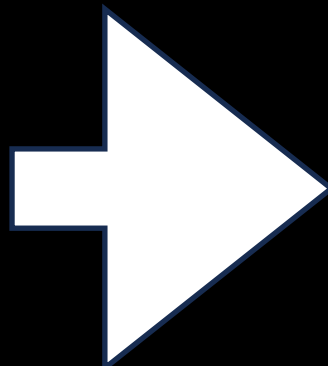


До какого числа мне нужно будет внести 2й платеж рассрочки ?



Здравствуйте, Евгений Александрович. Согласно информации, ближайший платеж по кредиту наличными должен быть осуществлен до 17 апреля 2025 года. Следующий платеж нужно внести до 17 мая 2025 года.

Название поля	Значение
Дата платежа	17.04.2025
Проценты платежа	500
Размер долга	2521
Размер платежа	3021
...	...
Поле № 300	XXX

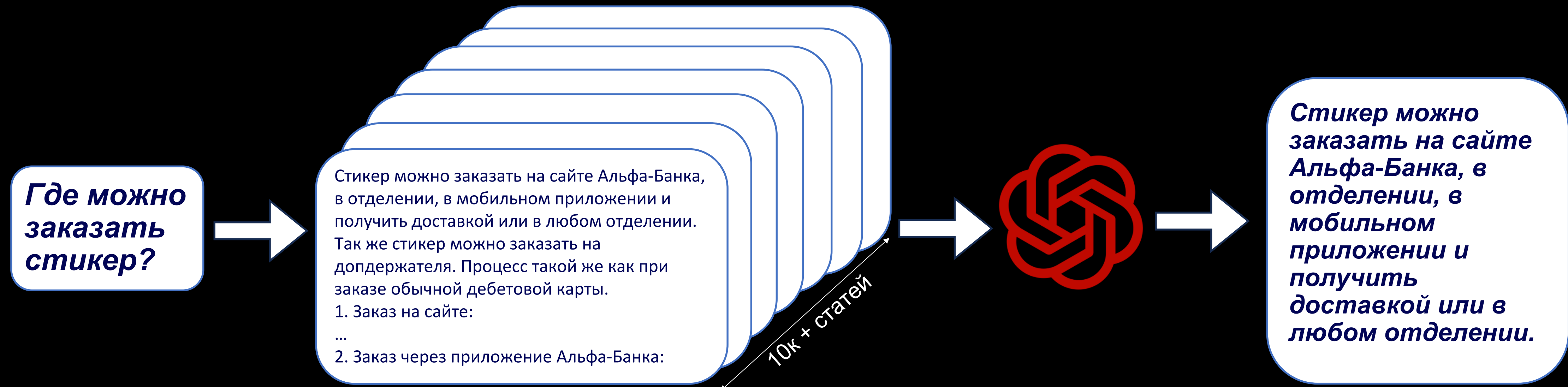


01 Берем 300 параметров клиента

02 Переводим таблицу в Markdown

03 Подаем в LLM с вопросом клиента

СХЕМА RAG-ПАЙПЛАЙНА



01. КЛИЕНТ
ЗАДАЕТ ВОПРОС

02. МОДЕЛЬ НАХОДИТ
РЕЛЕВАНТНЫЕ СТАТЬИ

03. LLM отвечает на вопрос клиента, опираясь
на информацию из найденных статей

! УСПЕХ RAG НА 90% ЗАВИСИТ ОТ РЕАЛИЗАЦИИ RETREIVE-ЧАСТИ И ПОЛНОТЫ БАЗЫ ЗНАНИЙ

СРАВНЕНИЕ LLM В RAG-ПАЙПЛАЙНЕ

Название метрики	LLM 7B	LLM 32B	LLM 70B	LLM 671B
Правдивость	3.0	3.5	3.9	3.5
Релевантность	2.9	3.8	3.4	3.5
Скорость	2	3	5	30

01. ВЫБОР LLM ВЛИЯЕТ НА
КАЧЕСТВО НЕ БОЛЕЕ, ЧЕМ НА 20%

02. КАЧЕСТВО МОЖНО НАРАСТИТЬ
ЗА СЧЕТ УВЕЛИЧЕНИЯ ЧИСЛА
ПАРАМЕТРОВ LLM

03. СКОРОСТЬ ОТВЕТА НЕОБХОДИМО
УЧИТЫВАТЬ ПРИ ВЫБОРЕ LLM

МОДУЛЬ ДОСПРАШИВАНИЯ



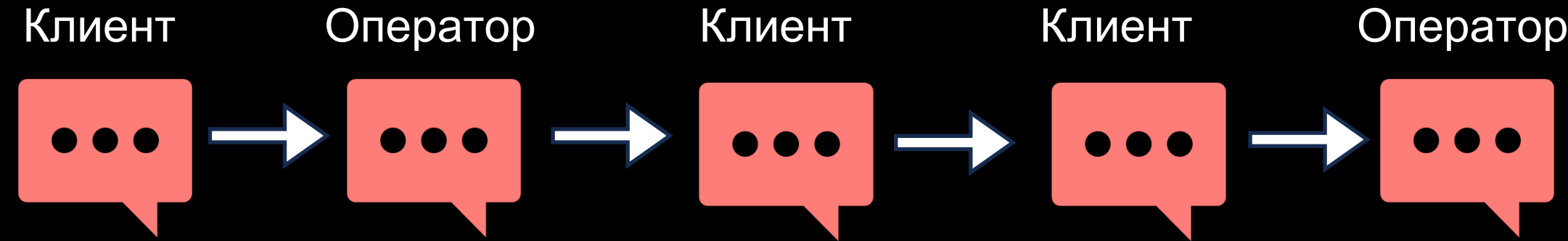
! Модуль получает контекст из описаний максимально-релеватных интенгов

КЛАССИФИКАТОР ЦЕЛЕВЫХ ДЕЙСТВИЙ



Признаковое описание

Контекст диалога

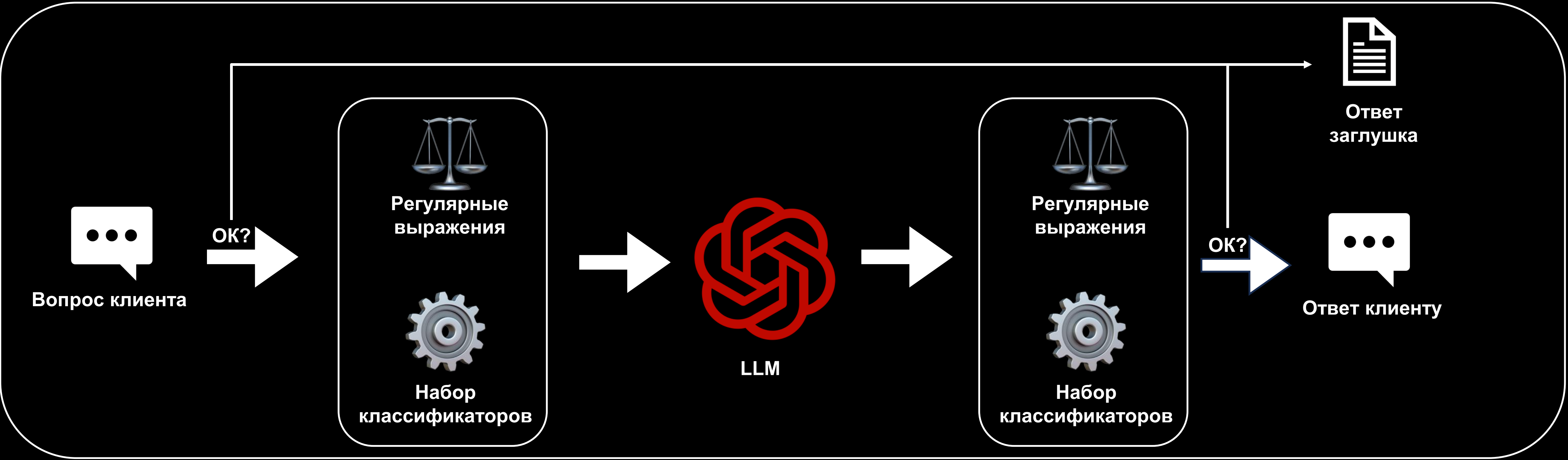


Целевая переменная



Класс целевого действия в интерфейсе

МОДУЛЬ ЦЕНЗУРИРОВАНИЯ



Юридический и репутационный риск

- Политические тематики
- Вопросы про конкурентов
- Экстремизм
- ...

Финансовый риск

- Мошенничество
- Утечка персональных данных
- Невыгодные рекомендации
- ...

Социально-чувствительные тематики

- Ненормативная лексика
- Эмоциональный негатив
- Религия
- ...

ВЕКТОРНОЕ КЭШИРОВАНИЕ ЗАПРОСОВ-ОТВЕТОВ

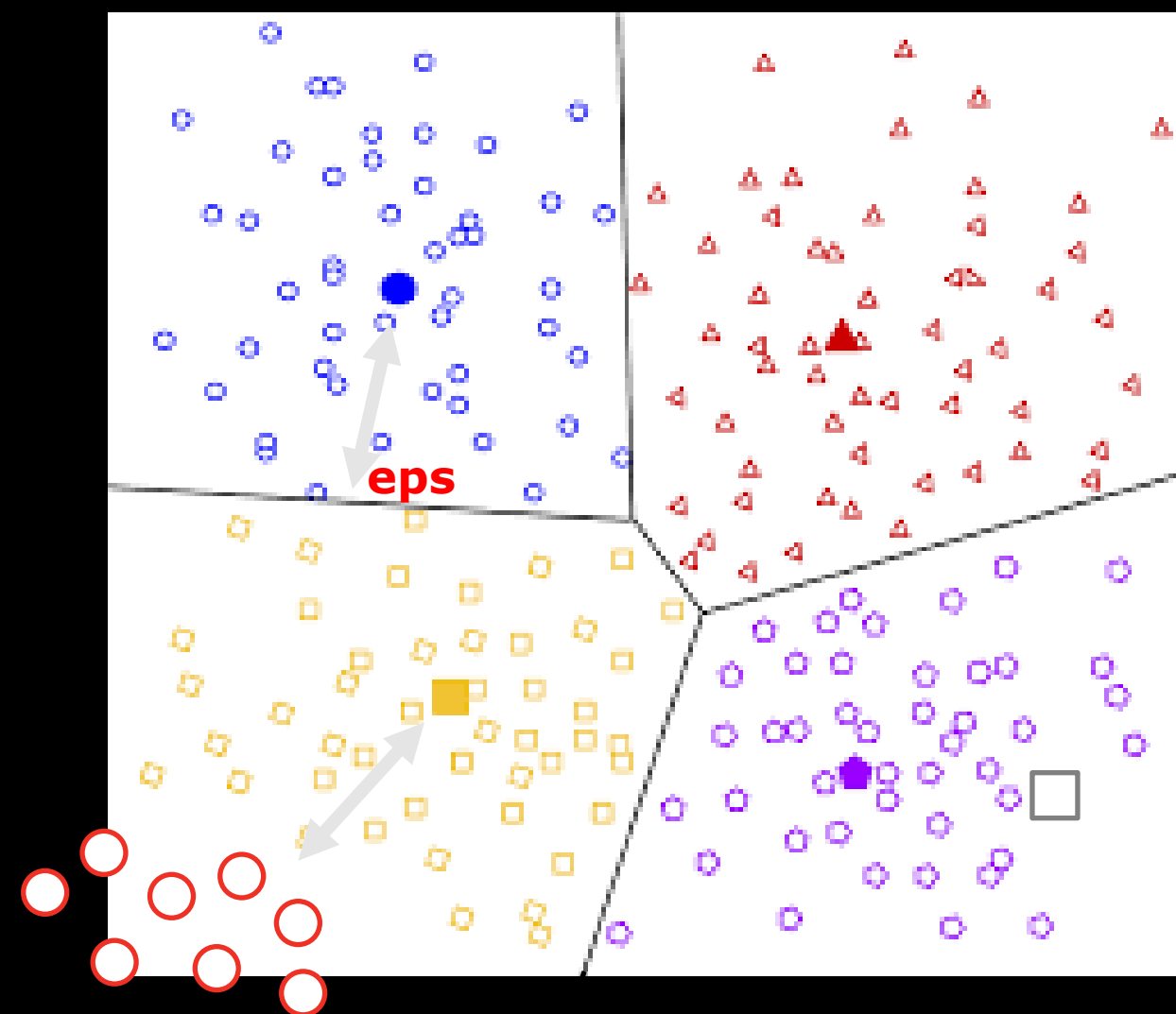
Наполнение

1. Накапливаем базу вопросов
2. Векторизуем данные
3. Выделяем плотное подмножество
4. Генерируем ответы на вопросы LLM

Обновление

1. Добавить дельту (расстояние больше ϵ) в базу вопросов
2. Обновить плотное подмножество
3. Генерируем ответы на вопросы LLM

“Не пришел кэшбек” = “не пришел Кэшбек»
= “не пришел cash back”...



Применение

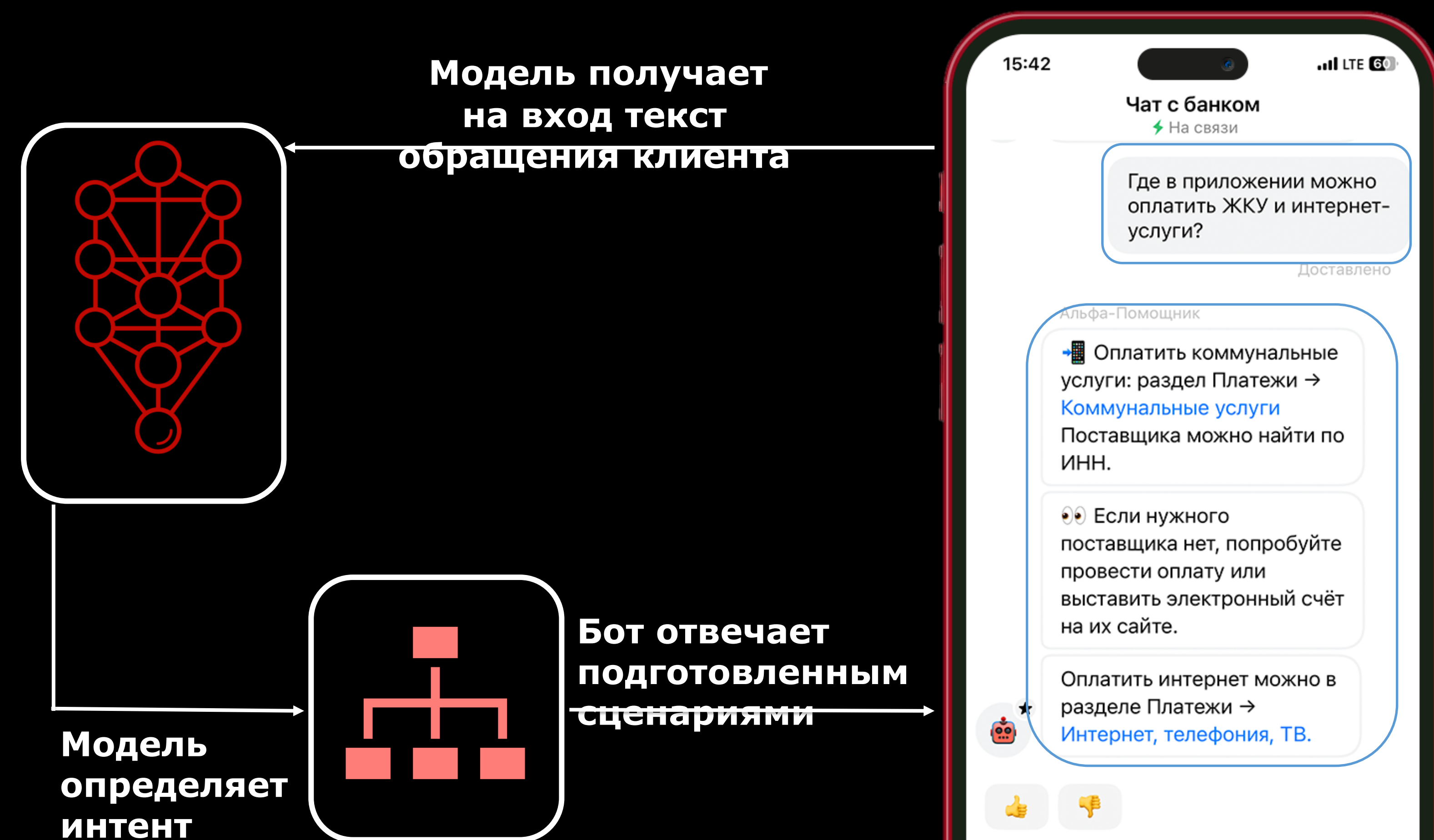
1. Считаем расстояние от вектора запроса до ближайшего в кэше
2. Используем ответ из кэша, если расстояние меньше ϵ

! УМНЫЙ КЭШ ПОЗВОЛЯЕТ ЗАМЕНИТЬ ДО 70% ОБРАЩЕНИЙ К LLM

ОПТИМИЗАЦИЯ ЧАТ-БОТА 1.0

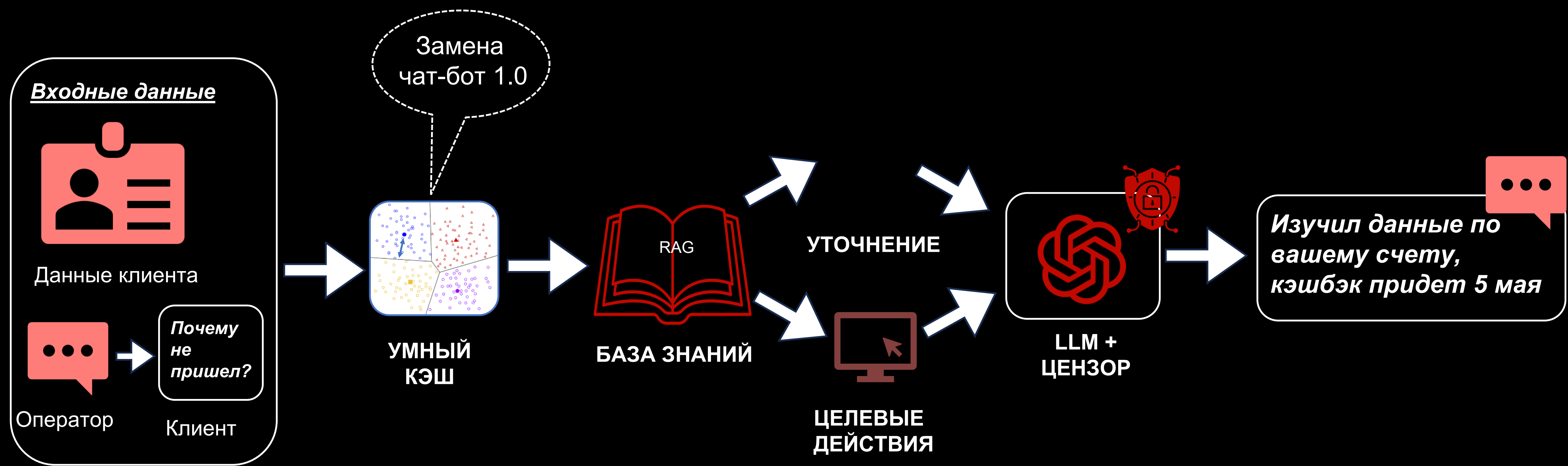
01. ЧАТ-БОТ 1.0 СПРАВЛЯЕТСЯ С ОБСЛУЖИВАНИЕМ КЛИЕНТОВ ПРИ ПОМОЩИ ПАРЫ ТЫСЯЧ ПРЕДЗАПИСАННЫХ ОТВЕТОВ

02. УМНЫЙ КЭШ СПОСОБЕН ПО КАЧЕСТВУ ЗАМЕНИТЬ ПРЕДЗАПИСАННЫЕ СЦЕНАРИИ



! ВЕРИФИЦИРОВАННЫЙ УМНЫЙ КЭШ ПОЗВОЛИТ ОПТИМИЗИРОВАТЬ ПРОЦЕСС СОЗДАНИЯ СЦЕНАРИЕВ

ЧАТ-БОТ 2.0



! ВЕРИФИЦИРОВАННЫЙ УМНЫЙ КЭШ ПОЗВОЛИТ ОПТИМИЗИРОВАТЬ ПРОЦЕСС СОЗДАНИЯ СЦЕНАРИЕВ

ПРИНЯТИЕ РЕШЕНИЙ НА ОСНОВЕ ДАННЫХ

KARPOV.COURSES



Какие типы задач можно и нельзя решать при помощи машинного обучения?



Какое требуется вовлечение владельца бизнес процесса?



Как устроен жизненный цикл моделей машинного обучения?



Зачем проводить аб-тесты и мониторить качество работы внедренных моделей?



ОСНОВНЫЕ ВЫВОДЫ

- 1. Уровень развитие ИИ позволяет обслуживать клиентов на уровне оператора**
- 2. Имитация оператора возможна за счет интеграции данных о клиенте, реализации механизмов доспрашивания и rag-пайплайна, а также классификатора целевых действий оператора**
- 3. Переиспользование интеграции данных в интерфейса оператора позволяет существенно ускорить процесс внедрения технологии**
- 4. Качество работы rag-пайплайна на 90% зависит от качества реализации retrieve-части**
- 5. LLM способны в сыром виде работать с табличными данными**
- 6. «Умный кэш» позволяет сэкономить до 70% вызовов LLM и заменить Чат-бота 1.0**
- 7. Переиспользование классификатора интенгов позволит повысить качество доспрашивания**
- 8. Модуль цензурирования на основе классификаторов и регулярных выражений защитит ваше решение от атак и галлюцинаций**

Оставайтесь на связи



Нескучный Data Science
10000+ подписчиков

