

# LLM в анализе клиентского опыта

Детализация клиентских  
сообщений и оценка качества  
работы моделей

Олег Зельдин

Апекс Берг

# Классификация текстовых отзывов от клиентов (обратная связь, жалобы)

# 8 161 отзыв о налоговой службе

Один год • два языка

6 515 английских • 1 646 французских

13 заранее заданных критериев  
качества сервиса

Сроки • ясность • доступность



Один отзыв может затрагивать несколько критериев

# 13 категорий качества обслуживания

Заранее заданные категории для анализа отзывов



Возможность  
эскалации



Доступность  
для людей  
с ограничениями



Наличие и  
доступность  
обслуживания



Ясность  
информации



Полнота  
информации



Согласованность



Удобство



Возможность  
найти информацию



Точность  
информации



Формат  
информации



Обслуживание  
на официальных  
языках



Профессионализм  
сотрудников



Своевременность  
обслуживания

# Тема и основание в тексте

Учебный пример

«Жду ответа месяц»



сроки

«Инструкция непонятна»



ясность

«Раздел сайта не могу найти»



поиск  
информации



Релевантность теме  $\neq$  удовлетворённость

# Динамика упоминаний по группам клиентов

## Сравниваем ранний и поздний периоды

Ранний период ● —————> ● Поздний период

| Группа клиентов                                                                                 | Критерий качества                                                                             | Ранний период                                                                                               | Поздний период                                                                                             |
|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
|  Возраст 60+   |  Ясность     | <br>Связь не обнаружена  | <br>Связь обнаружена    |
|  Возраст ≤ 19  |  Доступность | <br>Связь обнаружена     | <br>Связь не обнаружена |
|  Англоязычные |  Сроки      | <br>Связь не обнаружена | <br>Связь обнаружена   |

Иллюстративные примеры авторов

**В фокусе — динамика частоты упоминаний  
критериев качества по группам клиентов**

# Совпадение с категориями организации

**24,27%**

**Zephyr-7B-β**  
с инструкциями

**66,64%**

**Mistral-7B-Instruct-v0.2**  
с адаптацией

Авторский показатель similarity



**Одновременно  
изменились модель  
и подготовка**



**Это не полная проверка  
качества классификации**

# Как рассчитывается Similarity

Совпадение категорий модели и сотрудников

$$\text{Similarity (\%)} = \frac{\text{Число отзывов с совпавшей категорией}}{\text{Общее число отзывов}} \times 100$$

Тему модели сравнивают с категорией, назначенной сотрудником.

Условный пример



Правило совпадения при нескольких темах описано неполностью.

# Предложение по развитию методики

# Пробел в описании оценки качества

Что не раскрыто в рассмотренном кейсе

Учебный  
пример



Эталон: Сроки + Ясность

| Критерии модели         | Верно | Пропущено | Лишнее |
|-------------------------|-------|-----------|--------|
| Только Сроки            | 1     | 1         | 0      |
| Сроки + Профессионализм | 1     | 1         | 1      |
| Сроки + Ясность         | 2     | 0         | 0      |
| Все 13 критериев        | 2     | 0         | 11     |



Общее совпадение есть. Ошибки разные.

В статье недостаточно ясно описан учёт частичных совпадений, пропусков и лишних назначений.

# Шаг 1. Устанавливаем правила сравнения

Предлагаемый метод • Учебный пример

Один отзыв → полный набор критериев



**Эталон:**  
Сроки + Ясность



**Модель:**  
Сроки + Профессионализм



**Сроки**  
Найдено правильно



**Ясность**  
Пропущено



**Профессионализм**  
Назначено лишним



**Каждый критерий учитываем один раз на отзыв**

Заранее фиксируем определения критериев и правила их назначения

# Шаг 1. Готовим эталон и матрицу весов

Предлагаемый метод



Два эксперта  
независимо



Все  
обоснованные  
критерии



Разбор  
расхождений



Полный  
экспертный  
эталон



Матрица весов  
ошибок

Для каждого критерия:

Вес пропуска



Вес лишнего  
назначения



Каждый вес — от 0 до 1



Эталон и веса фиксируем до проверки модели

# Шаг 1. Пример матрицы весов

Предлагаемый метод • Учебный пример

| Критерий качества обслуживания | Вес пропуска | Вес лишнего назначения |
|--------------------------------|--------------|------------------------|
| Сроки                          | 0,8          | 0,3                    |
| Ясность                        | 0,8          | 0,3                    |
| Точность                       | 1,0          | 0,5                    |
| Полнота                        | 0,4          | 0,2                    |
| Профессионализм                | 0,6          | 0,2                    |

0 — без штрафа • 1 — максимальный вес ошибки

Весы задают эксперты до теста. Одна матрица для всех моделей.

Показан фрагмент матрицы; значения учебные.

# Шаг 2. Независимо сравниваем модели

Предлагаемый метод



Обучение



Настройка



Независимый тест

**Тест не используем для настройки**



Одна контрольная  
выборка



**Mistral-7B-Instruct-v0.2**



До адаптации



После адаптации

Один эталон • Одна матрица весов • Одни правила

**Так проверяем эффект именно адаптации**

# Шаг 3. Считаем оценку одного отзыва

Предлагаемый метод • Учебный пример

$$Q = \frac{C}{C + \sum v_i + \sum u_i}$$

$C$  — число верных критериев  
 $v_i$  — вес пропуска критерия  
 $u_i$  — вес лишнего назначения критерия  
Суммы — только по фактическим ошибкам

**Эталон:** Сроки, Ясность, Точность, Полнота  
**Модель:** Сроки, Точность, Профессионализм  
**Верно:** Сроки и Точность →  $C = 2$ .

| Ошибка                 | Штраф |
|------------------------|-------|
| Пропущена Ясность      | 0,8   |
| Пропущена Полнота      | 0,4   |
| Лишний Профессионализм | 0,2   |

$$Q = 2 / (2 + 0,8 + 0,4 + 0,2) \approx 0,588$$

Диапазон  $Q$ : от 0 до 1 при ненулевом знаменателе.

58,8 из 100

# Шаг 4. Проверяем, где модель ошибается

Предлагаемый метод

## По критериям качества



Сроки



Ясность



Точность

Считаем частоты разных  
типов ошибок по каждому  
критерию

## По группам клиентов



Язык



Возраст



Пол

Для всех срезов сохраняем  
те же правила и веса

# Что получаем в результате проверки

Предлагаемый метод



**Понятный  
эталон**

Полный набор  
критериев



**Прозрачные  
ошибки**

Что пропущено  
и что назначено  
лишним



**Одна оценка  
отзыва**

С учётом значимости  
каждой ошибки



**Основа  
для решения**

Где применять  
модель, а где нужна  
проверка



**В отчёте: эталон → ошибки → веса → оценка**



**Проверяем, насколько надёжно модель выявляет проблемы  
сервиса в клиентских отзывах**

# Проблема 2. Не раскрыт порог назначения критерия

Ограничение описания в статье

**Модель:**  
релевантность  
от 1 до 5



Какой балл считать  
достаточным?



Критерий  
присутствует:  
1 или 0



**Пример: «Сроки» — балл 3**  
**Назначить критерий или нет?**



**Высокий балл —**  
критерий явно затронут



**Низкий балл —**  
мало оснований для назначения



**Переход от балла к решению в статье не описан**  
Балл 3 приведён как учебный пример.

# Решение. Дробная шкала и проверяемый порог

Предлагаемый метод • Учебный пример

- 1** Модель оценивает релевантность



Шкала от 0 до 1;  
например, с шагом 0,01

- 2** Выбираем порог для каждого критерия



По экспертному эталону  
на настроечной выборке

- 3** Фиксируем пороги до независимого теста



Критерий: Сроки •  
Учебный порог  $t = 0,70$

$r = 0,68 \rightarrow$  не назначаем

$r = 0,82 \rightarrow$  назначаем



**Правило:**  $r \geq t \rightarrow$   
критерий присутствует



Частоты пропусков и лишних  
назначений проверяем  
для каждого порога

Дробная шкала сама по себе  
не повышает точность

# Развитие методики. Оцениваем негатив по каждому критерию

Предложение Олега • Учебный пример



## Релевантность

Затронут ли критерий  
в отзыве?



## Уровень негатива

Насколько выражено  
недовольство?

От 0 — не выражено до 1 — сильно выражено

“ Ответили быстро, но объяснение совершенно непонятное. ”

| Критерий | Релевантность | Негатив |
|----------|---------------|---------|
| Сроки    | 0,95          | 0,00    |
| Ясность  | 0,95          | 0,90    |



## Акцент — на критериях с выраженным негативом

Баллы учебные; это развитие методики, а не результат статьи

# Оценка клиентского опыта по разговорам/чатам

# 8 880 разговоров: 537 с клиентской оценкой

30 дней • один суббренд

Автоматический чат и звонки



**537**

сессий с оценкой клиента — около 6%

**432** — общий балл в чате

**105** — связаны с расширенной анкетой

Анкета: 92 кейса, связанные со 105 сессиями



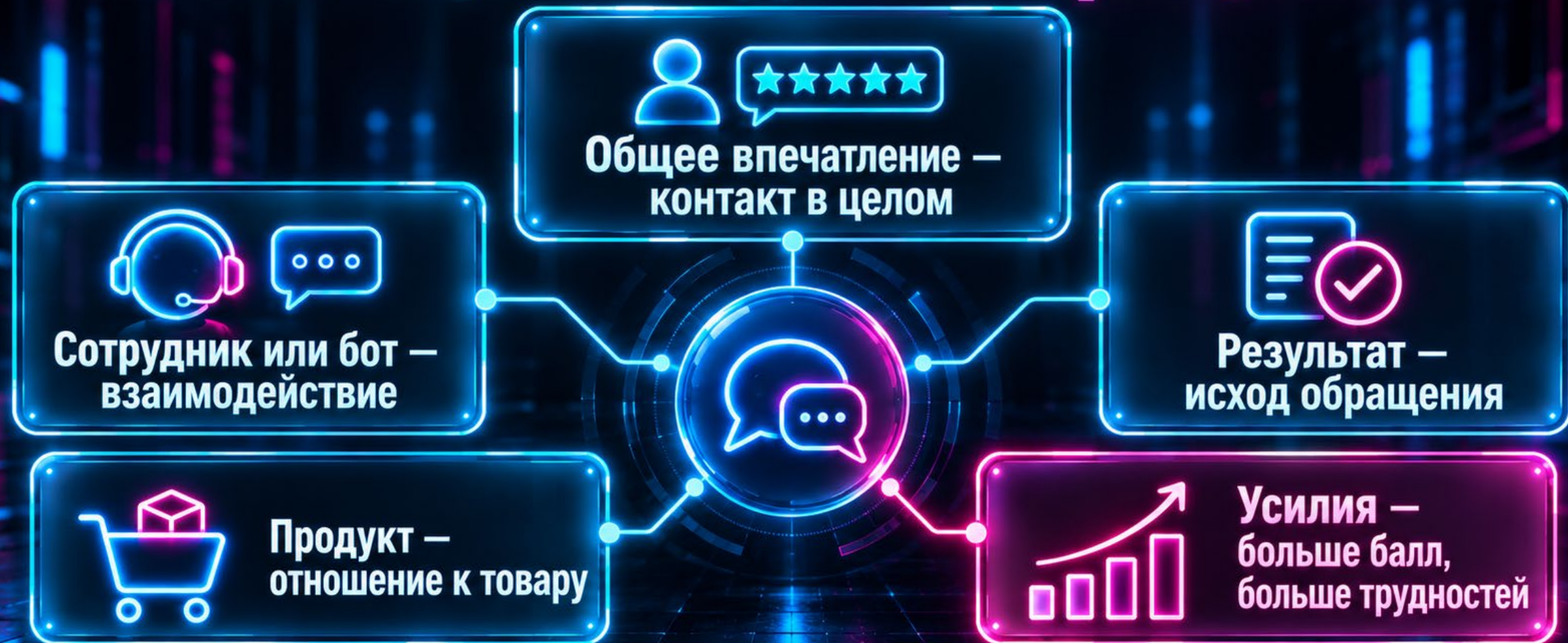
**8 343**

сессии без оценки  
клиента



LLM анализирует тексты обеих групп

# Пять оценок одного обращения



**Шкала 1–5 • мало сведений → без оценки**

# Работа сотрудника и впечатление от продукта



**427**

**из 8 880:**  
сотрудник высоко,  
продукт низко



**4,8%** корпуса  
• оценки LLM

Пример: протекающая упаковка



Сотрудник

**5**



продукт

**2**



общий

**4**



Все три балла — оценки модели

# Что показала проверка оценок



Чат: **385 пар**  
корреляция **0,65**



Отдельный опрос:  
**81 случай**

**59%**

— точное совпадение

**94%**

— разница не более 1 балла



**Корреляция — не процент точности**

# Отвeтившие и неответившие на опрос

Шкала 1–5



С ответом  
на опрос —

**3,62**



Без ответа  
на опрос —

**2,91**



**Обе оценки рассчитала LLM**



Оценку неответивших  
независимо не проверили

# Применение в КЦ. Рекомендации

# Оценка опыта и анализ тем вместе

Оценка контакта → тема → управленческий вопрос



К чему относится  
недовольство?



Какие проблемы  
повторяются?



Где нужен  
дополнительный  
разбор?



Наше предложение применения **подходов двух статей**

# Как улучшить методику оценки разговоров

## Предлагаемое развитие

1



### Совместная работа с экспертами

Проверять каждую составляющую опыта и выраженность негатива

2



### Учёт разных ошибок

Учитывать пропуски, лишние назначения и расхождения в баллах

3



### Проверка обеих групп

Отдельно проверять ответивших и неответивших на опрос

4



### Фактическое завершение задачи

Подтверждать результат данными CRM

5



### Устойчивость оценок

Проверять повторяемость и влияние инструкций модели

6



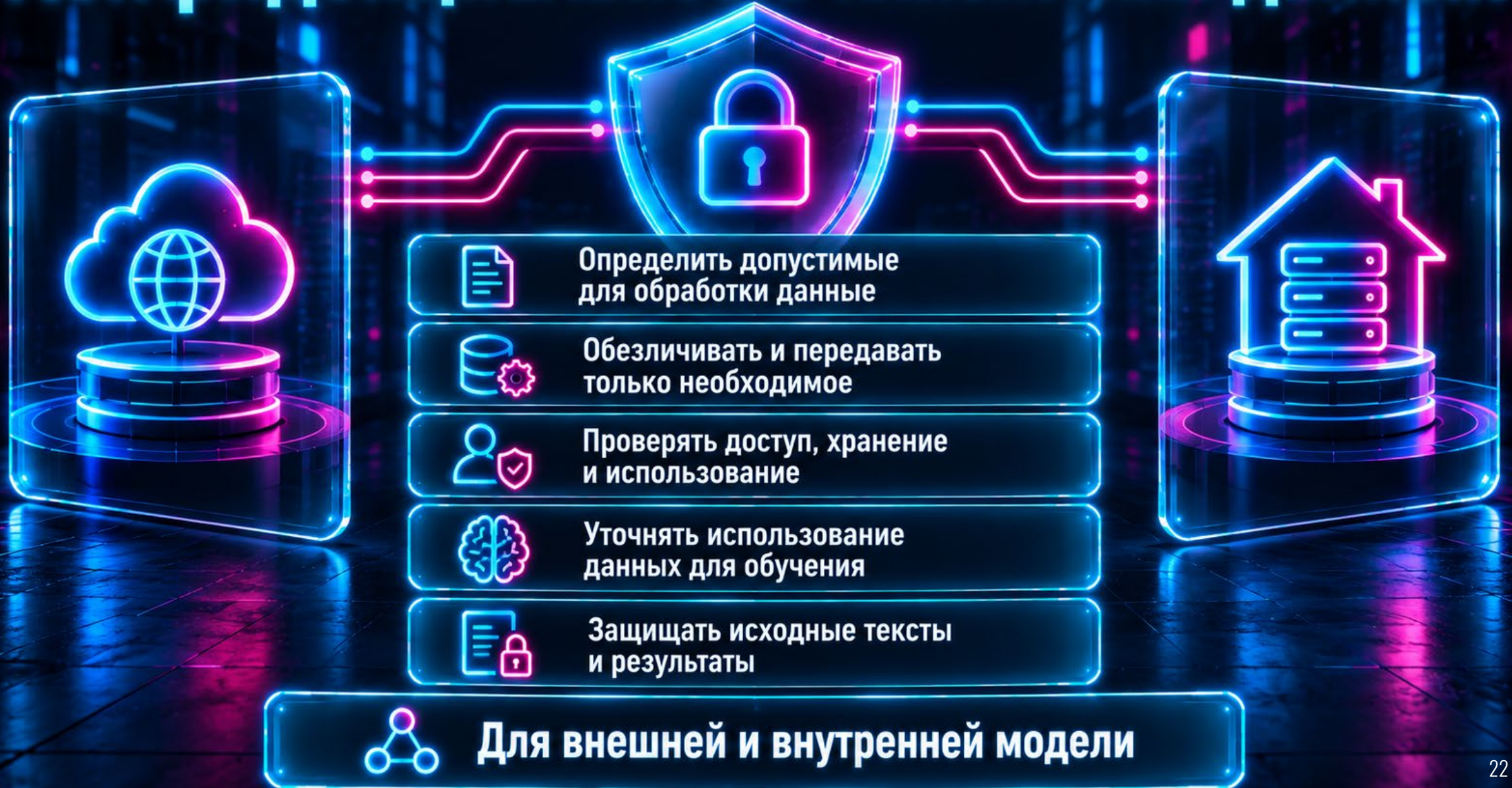
### Связь с бизнес-результатами

Сопоставлять оценки с повторными обращениями и поведением клиентов



**Цель — надёжно выявлять проблемы клиентского опыта**

# Конфиденциальность клиентских данных



# Источники вдохновения

Dimensionality in Satisfaction Ratings

<https://arxiv.org/abs/2607.11026>

LLM-based Models for Detecting Emerging Topics in Service Feedback

<https://arxiv.org/abs/2606.26595>

# СПАСИБО ЗА ВНИМАНИЕ!

---



## JOB QUALITY 2027

ХII МЕЖДУНАРОДНОЕ ИССЛЕДОВАНИЕ  
УДОВЛЕТВОРЕННОСТИ ОПЕРАТОРОВ  
КОНТАКТ-ЦЕНТРОВ

