Зайнулла Жумаев

к.т.н., старший разработчик-исследователь

ГК «ИнфоВотч»

Обеспечение доверия к генеративному ИИ в DLP-системе

INFOWATCH.RU

Генерация промежуточных верифицируемых методов

КЛАССИФИКАЦИЯ ТЕКСТОВ ЗАВИСИТ ОТ ЦЕЛЕВЫХ КАТЕГОРИЙ

INFOWATCH

Договор на проведение геолого-разведывательных работ

«Договор»

продление срока | рекламация | приложение к контракту

«Геологоразведка»

бурение | скважина | месторождение нефти

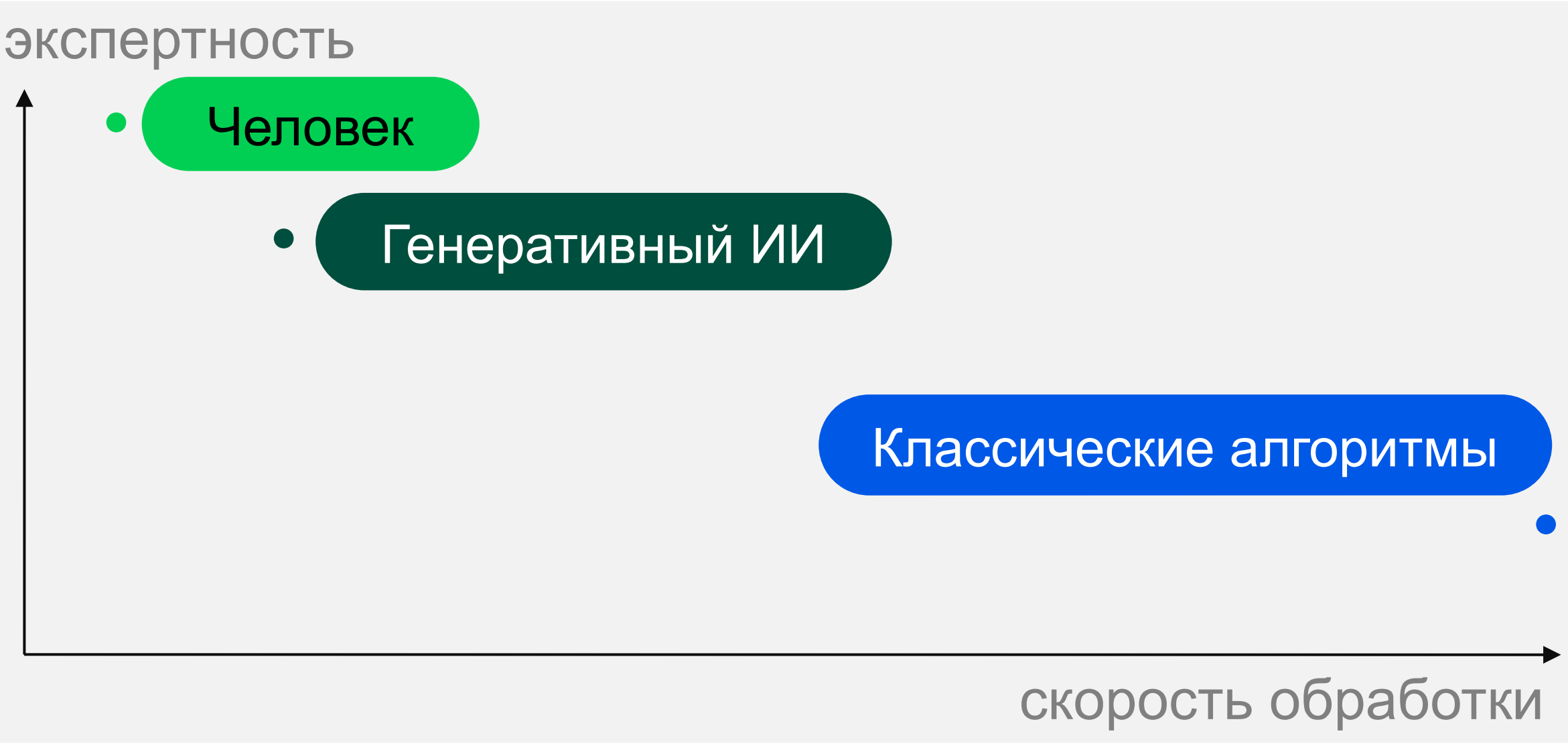
! Методы «мешка слов» работают плохо, важны ключевые словосочетания.

LLM В ПОТОК – ЭТО ДОРОГО И НЕ ВСЕГДА ТОЧНО

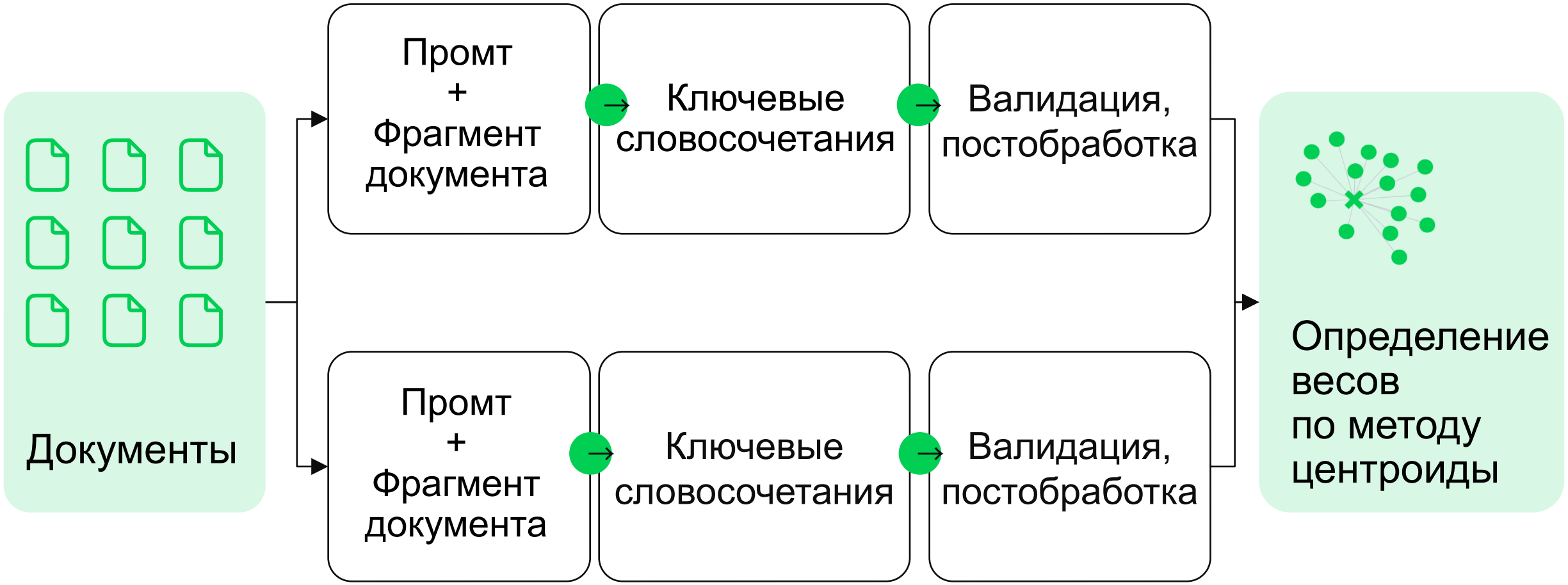
LLM хорошо классифицирует тексты, но требует кластеров GPU и работает на порядки медленнее классификации по словарям.

Классификация по словарям работает быстро и на CPU, но требует ручного труда экспертов-лингвистов.

ГЕНЕРАТИВНЫЙ ИИ МОЖЕТ НАСТРАИВАТЬ КЛАССИЧЕСКИЕ АЛГОРИТМЫ



ШТУЧНОЕ ПРИМЕНЕНИЕ LLM ДЛЯ ФОРМИРОВАНИЯ СЛОВАРЕЙ



ПРОМПТ УЧИТЫВАЕТ НАЗВАНИЕ КАТЕГОРИИ

INFOWATCH

Ты — инструмент
для извлечения ключевых
слов и словосочетаний
(n−грамм), характерных
для конкретной тематики.

Тематика: **{doc_category}**

Фрагмент документа: **{doc_chunk}**

Задача:

Проанализируй фрагмент текста
и выдели характерные слова,
фразы и словосочетания,
которые являются маркерами
темы "**{doc_category}**"

ИЗВЛЕКАЮТСЯ СЛОВОСОЧЕТАНИЯ

INFOWATCH

Категория Информационная безопасность

Текст термина	Вес	Хар.	Язык	Морф.	Регистр	Дли...	Обнаружен
Введите текст						=	<=
шифрование	10	✗	rus	✓	✗	1	119
требования ФСБ россии	7	✗	rus	✓	✗	3	141
государственная тайна	5	✗	rus	✓	✗	2	133
запрос ЕГРЮЛ	4	✗	rus	✓	✗	2	20
орган ФНС	4	✗	rus	✓	✗	2	19
папка	4	✗	rus	✓	✗	1	20
ключевая информация	4	✗	rus	✓	✗	2	54
аппаратная платформа	4	✗	rus	✓	✗	2	5
электронная цифровая подпись	4	✗	rus	✓	✗	3	105

Учитывается фактическое количество обнаружений термина

ПОЛУЧАЕМ ВЕРИФИЦИРУЕМУЮ ТЕХНОЛОГИЮ

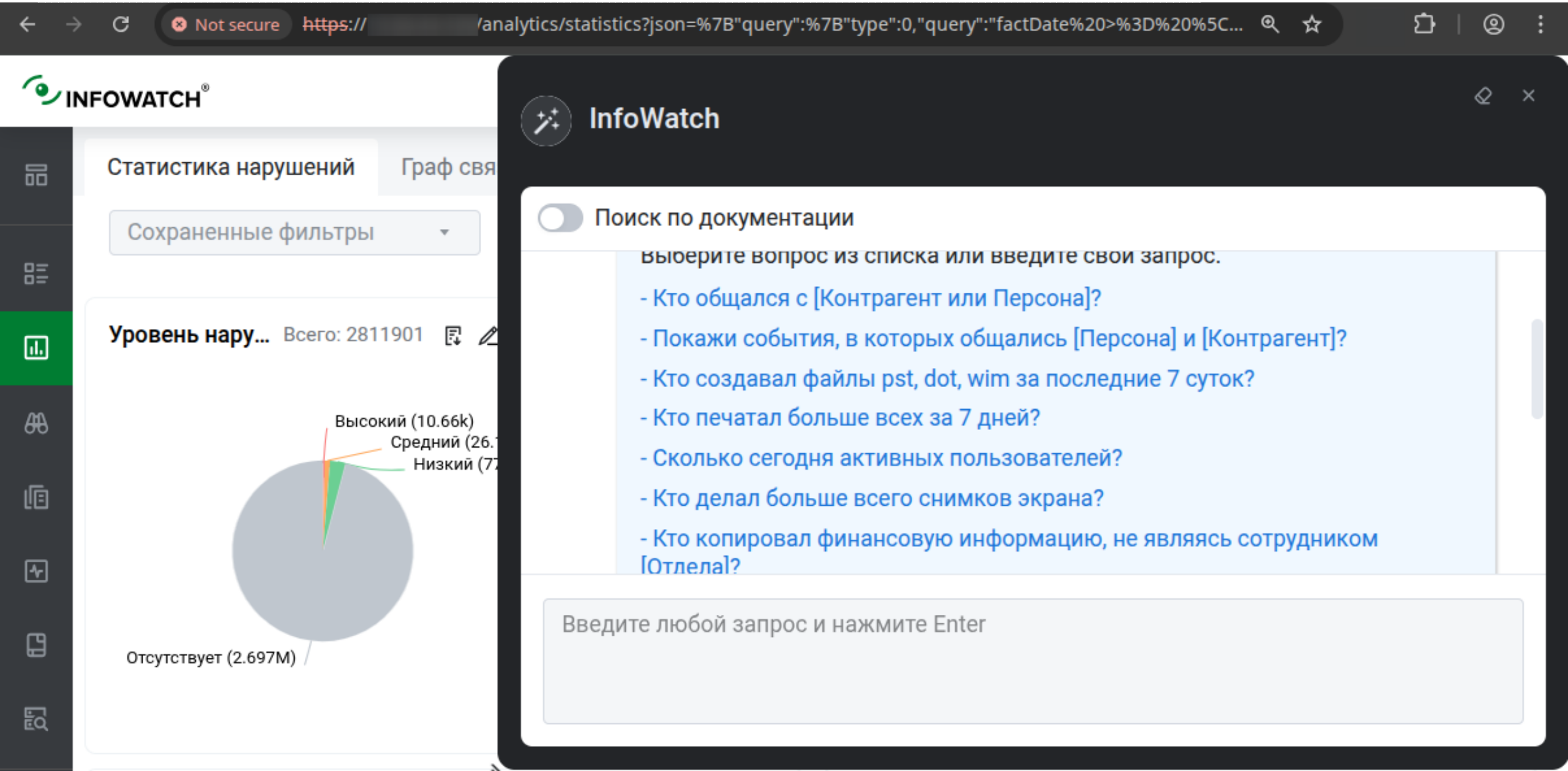
INFOWATCH

- ЛПС (ложные термины) исключаются подсчётом.
- Толерантность к небольшой доле ЛОС за счёт большого количества терминов.
- Общие термины удаляются (вектора категорий ортогональны) – меньше ЛПС в потоке.

Контроль
на тестовых данных

ДОБАВЛЕНО ВЗАИМОДЕЙСТВИЕ С DLP ЧЕРЕЗ ЧАТ

INFOWATCH



ПРИМЕР ОЖИДАЕМОГО РЕЗУЛЬТАТА

INFOWATCH

«Найди события, с 24 января по 26 февраля 2023, у которых было нарушение политики Персональные данные, с типом файла Архивы, с типом события Файловые операции и отсутствовал уровень нарушения»

Внутреннее представление запроса

```
SELECT * FROM events
WHERE factDate BETWEEN ('24.01.2023', '26.02.2023')
    AND policies IN ('Персональные данные')
    AND attachmentsMimesSplit IN ('Архив')
    AND factTypeSplit IN ('Файловые операции')
    AND violationType IN ('No');
```

СГЕНЕРИРОВАННЫЙ LLM DSL-ЗАПРОС ПОСТОБРАБАТЫВАЕТСЯ

INFOWATCH

Системный промт + Запрос пользователя → Извлечение фрагментов DSL-запроса с помощью регулярных выражений → Валидация фрагментов DSL-запроса → Формирование DSL-запроса

```
SELECT * FROM events
WHERE factDate BETWEEN ('24.01.2023', '26.02.2023')
    AND policies IN ('Персональные данные')
    AND attachmentsMimesSplit IN ('Архив')
    AND factTypeSplit IN ('Файловые операции')
    AND violationType IN ('No');
```

МЕТРИКИ КАЧЕСТВА УЧИТЫВАЮТ ВАРИАТИВНОСТЬ ВЕРНЫХ ЗАПРОСОВ

INFOWATCH

Ожидание: **условие1** and **условие2** and **условие3**

Получено: **условие1** and **условие2** and **условие4**

$$\text{Качество (IoU)} = \frac{2}{2+1+1}$$

Контроль
на этапе выполнения

ИСПОЛЬЗОВАНИЕ СПЕЦИАЛИЗИРОВАННЫХ ДЛЯ ИИ И ОБЩИХ ОГРАНИЧИТЕЛЕЙ

INFOWATCH

- *Regex* и *Shema guided reasoning*
- Только подмножество *SELECT* запросов
- Ограничения по объёму данных на уровне фильтра (условие по датам)
- Ограничения через параметры *max_memory_usage* и *max_execution_time*

ИСПОЛЬЗОВАНИЕ СПЕЦИАЛИЗИРОВАННЫХ ДЛЯ ИИ И ОБЩИХ ОГРАНИЧИТЕЛЕЙ

INFOWATCH

01 Генерировать промежуточные верифицируемые классические технологии и автоматически корректировать

02 Контроль метрик качества на тестовой выборке

03 Ограничения на этапе выполнения

INFOWATCH

Спасибо за внимание!

Обеспечение доверия к генеративному ИИ в DLP-системе

ЗАЙНУЛЛА ЖУМАЕВ

к.т.н., старший разработчик-исследователь ГК «ИнфоВотч»

INFOWATCH.RU