Малые данные + генеративный ИИ

01 ноября 2023

Гуреев Дмитрий

2 слова о себе

Гуреев Дмитрий
(MBA, выпускник CDTO-3 Сколково, соавтор курса «ИИ для бизнеса», приглашенный эксперт по работе с данными в МШУ Сколково)

CDTO Биовитрум

- 18 лет в области лабораторной медицины (со стороны поставщиков)
- Карьерный путь от сервисного инженера до генерального директора
- Обширный операционный опыт, в т.ч. антикризисный: от продаж до R&D
- Цифра всегда помогала делать карьеру

Исследование:

- Управление малыми данными как сквозной навык
- Роль локальных LLM в построении бизнес – процессов нового типа

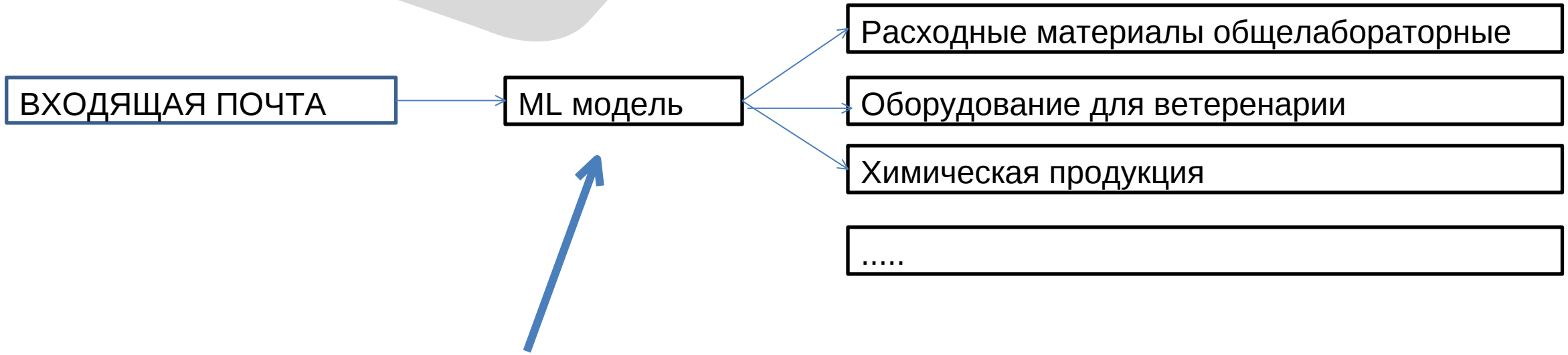
Малые данные есть у всех. Большие у единиц.

Малые данные (МД) - оперативные и справочные бизнес-данные, хранящиеся в информационных системах.

Что бизнес хочет от малых данных?

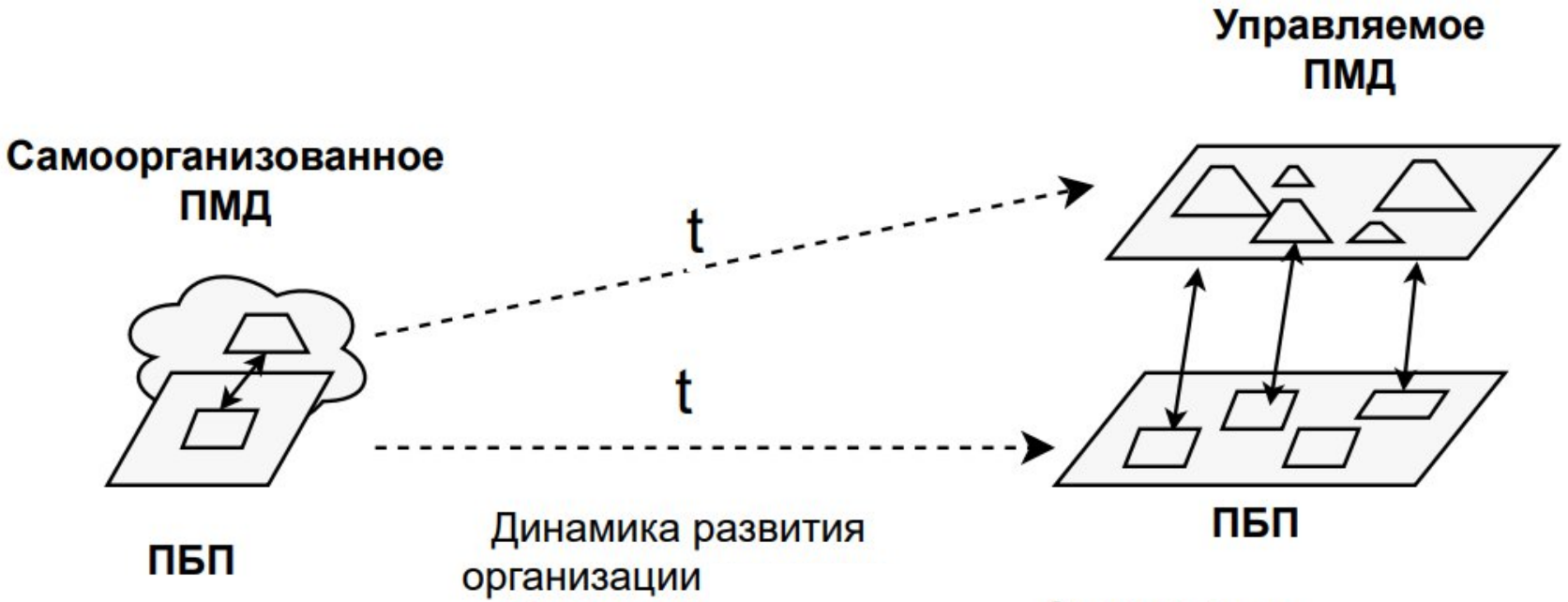
- Дашборды в один клик из всех систем и интеграция всего со всем. (интероперабельность)
- Чтобы не было ошибок в данных и в них можно верить. (валидность и консистентность)
- Доступ к данным был открытым внутри и закрытым извне. (ИБ)
- На их основе создавать микросервисы. (различные задачи автоматизации/цифровизации)

Классификация входящих запросов клиентов по товарным группам



Классификационная модель натренированная на справочнике товаров из 1С (малые данные)

Автоматизируя хаос, получим автоматизированный хаос



Концептуальные модели данных и прочее

Схемы бизнес процессов

Определения

- **Малые данные (МД)** - оперативные и справочные бизнес-данные, хранящиеся в информационных системах.
- **Пространство малых данных (ПМД)** - совокупность всех бизнес-данных, которые создаются, преобразовываются и хранятся в контролируемых организацией информационных системах.
- **Пространство бизнес-процессов (ПБП)** - совокупность функций (закупка, маркетинг, продажа, производство...), которые делает компания в рамках своего основного бизнес-процесса.

Большие компании создают отделы Data Office,
средние компании – в лучшем случае чистят справочник контактов.

Дорого, долго и сложно.
А можно как-то иначе?

LLM (LARGE LANGUAGE MODELS) CHATGPT, YAGPT, GIGACHAT, LLAMA2, MISTRAL.....

ЭТИ МОДЕЛИ ЭВОЛЮЦИОННО УЛУЧШИЛИ ПРИНЦИП КОММУНИКАЦИИ ЧЕЛОВЕК – МАШИНА, ДОВЕДЯ ЕГО ДО СТИЛЯ ЧЕЛОВЕК - ЧЕЛОВЕК

Оценка рынка в базовых единицах (кг,литры) через LLM

Дано (из базы тендеров):
текстовая строка с названием медицинского товара и возможно с указанием размерности упаковки

Необходимо:
дообогатить строку столбцом с указанием сколько базовых единиц товара

Парафин для гистологической заливки Гистомикс , 5кг - Сколько кг?

Формалин 15х25мл в кор. 5 кор/уп - Сколько литров?

Ограничение:
- типов единиц и типов товаров неограниченно
- все записи различны, структуры записей нет и не будет
- есть товары (и их большинство), по котором нет экспертизы

LLM справились одним промтом

Ты технический писатель. Определи сколько литров в ... (промпт сократил)

Формалин 15х25мл в кор. 5 кор/уп.

X GPT 3.5 : 3.75 л.

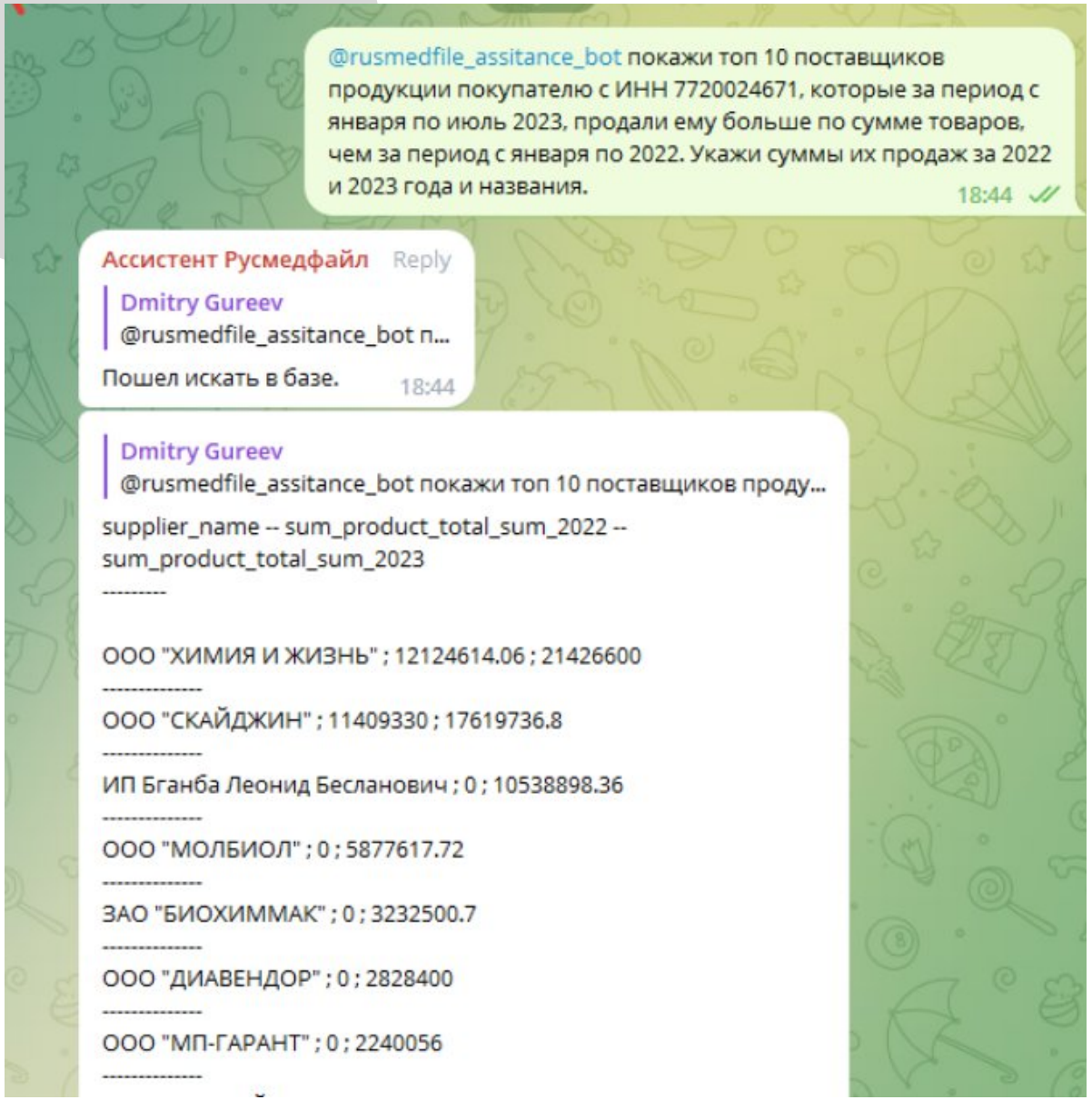
X YandexGPT: 5

V GPT4: 1.875

Профилирование данных в каталоге

- На входе примеры записей, на выходе тип данных (метаданные):
  —- Адрес
  —- Телефон
  —- .....

Text to SQL (лучший дашборд, которого нет)



Базовые знания о данных как объекте управления

+

LLM модели (+ Low Code)

=

Возможность управлять данными на эволюционно ином уровне (быстрее, дешевле, проще)

Основные вызовы

- Перенос LLM моделей в закрытый контур, так как соблазн использовать OpenAI – велик
- Удобный и быстрый Fine-Tune моделей
- Углубляющееся цифровое неравенство (знаний)

Время вопросов и ответов

Гуреев Дмитрий Владимирович
dvgureev@gmail.com
+ 7 (925) 741 94 53