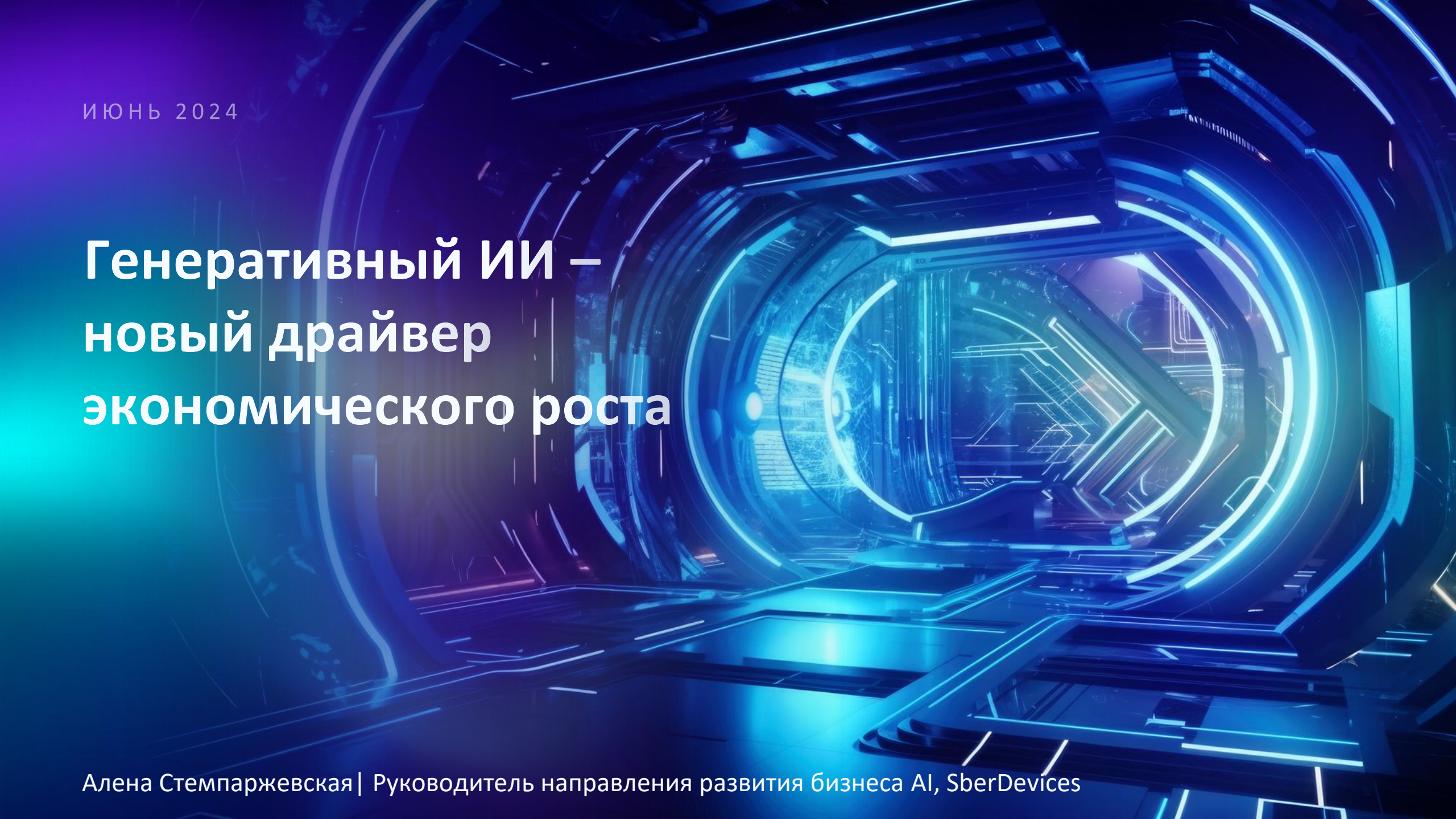


ИЮНЬ 2024

Генеративный ИИ – новый драйвер экономического роста

Алена Стемпаржевская| Руководитель направления развития бизнеса AI, SberDevices

SBER DEVICES

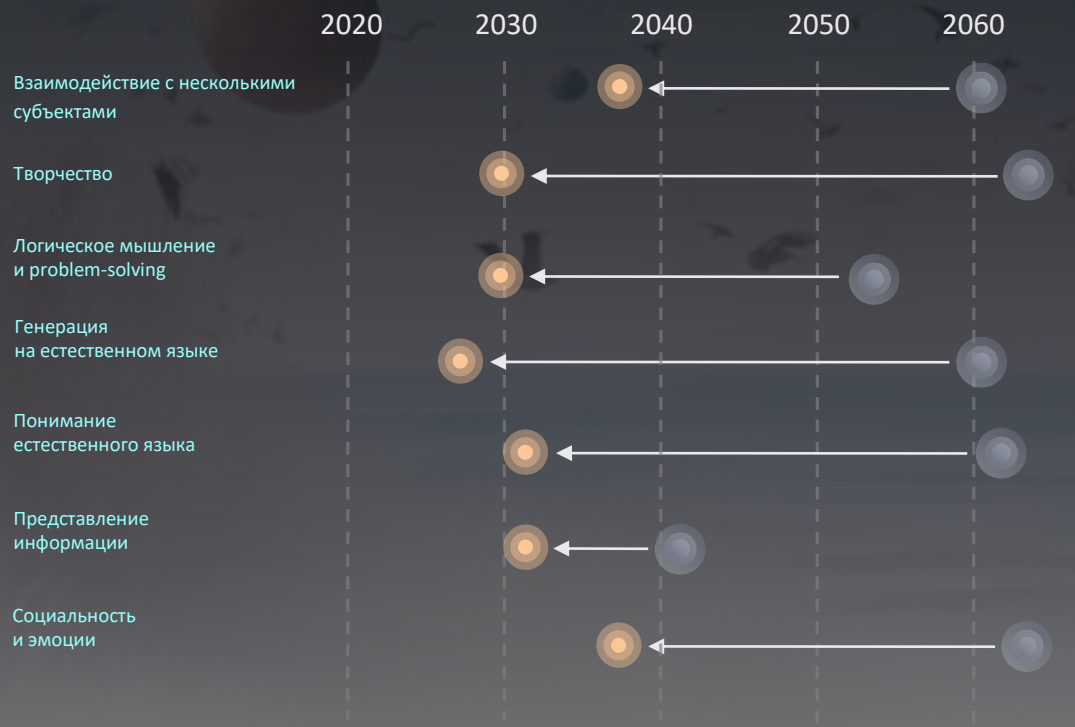
Ключевые моменты в развитии технологий

	Персональные компьютеры	Интернет	Мобильные приложения	AI
Период адаптации	20 лет	12 лет	6 лет	~ 3 года

LLM значительно приблизили выход AI на «человеческий» уровень

SBER DEVICES

Достижение AI уровня 25% топ-перформеров



2030

AI будет работать на среднем уровне человеческой производительности

2040

AI станет высокопроизводительным помощником человека

Источник: McKinsey

SBER DEVICES

SberDevices — технологический вендор в Сбере

Нейросетевая модель

GIGA CHAT

Цифровой офис

JAZZ by Sber

Телевизоры Sber
21 модель, 32” — 75”, OLED,
4K, Full HD,
HD Ready

Умные устройства

Умный дом

SberBoom &
SberBoom Mini

SberPortal

SberBox
Time

SberBox Top

SberBox

- Салют, GigaChat

от 8К контекст

= 12 страниц А4

GigaChat Lite(+) – 8/32к контекст – быстрая, для реализации базовых сценариев

GigaChat Pro – 8к контекст – для выполнения сложных инструкций и специальных доменных

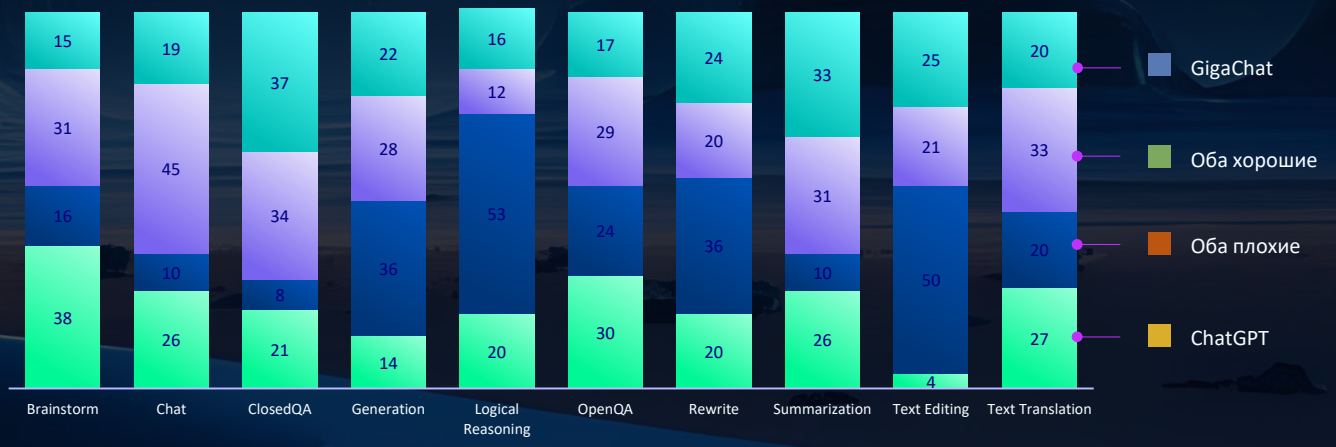
Embedder – для векторного представления текста

50 / 50

Side-by-Side
GigaChat-29B vs. ChatGPT-3.5

63,2

Massive Multitask Language Understanding (MMLU) Benchmark



7.5 ПБ

Это 47 Российских государственных библиотек

…или как если бы мы обработали всю литературу в мире 2,5 раза

1024

GPU

Для обучения модели

4 мес. можно было бы снабжать электричеством весь стадион Лужники

Варианты работы с GigaChat

Удобный пользовательский интерфейс

Общение с моделью в удобном формате в web-версии, Tg или VK

- Быстрая генерация идей
- Создание текстов
- Рисование картинок по текстовому описанию

Прямой запрос к AI модели

Для работы с базовой моделью «из коробки», при помощи инструкций

- Генерация
- Суммаризация
- Ответы на вопросы

Адаптация модели к бизнес-сценариям

Для «дообучения» модели данными компании и взаимодействия с внутренними системами

- Ответы по документации
- Чат-бот
- Суммаризация больших текстов

Готовое решение со встроенной AI моделью

Для решения конкретных задач бизнеса GigaChat уже интегрирован в ряд решений на рынке

- CRM
- RPA – платформа
- Офисное ПО
- Сервисы аналитики

Первый шаг

Составление инструкций

Создание сценариев

Готовые решения

Решения с уже интегрированной AI-моделью

GIGA CHAT inside

SaluteBot

Интеллектуальные чат-боты для обработки обращений клиентов и сотрудников без участия человека.

Обогащение реплик чат-бота, автоматизация диалогов

SaluteSpeech

Платформа для распознавания и синтеза речи.

Её технологии расшифровывают, озвучивают и превращают текст в аудиодорожку.

Генерация текста для озвучивания

SberJazz

Сервис для общения, работы и обучения.

Можно создавать видеовстречи любой продолжительности и переписываться во встроенном мессенджере.

Суммаризация итогов встречи

SaluteRPA & IDP

Программные роботы для автоматизации бизнес-процессов.

Помогут оптимизировать ресурсы сотрудников и сократить затраты на рутинные операции.

Диалог с документами в режиме чата

JIVO

Чат-платформа для связи с клиентами: чат на сайте, соцсети, мессенджеры и звонки.

Подсказки для оператора, быстрые ответы на сообщения пользователей

Какие задачи подходят для LLM уже сегодня

Исполнение инструкций

Выполнение поэтапных преобразований текстовых данных

Например:

— Перепиши текст […] в более формальном стиле

— Напиши сопроводительное письмо для рекламной кампании соблюдая структуру […]

— Составь ответ на обращение с учетом установленного регламента

Анализ данных

Например:

— Проанализируй текст и присвой тэги из базы

Структурирование информации

Помощь в написании запросов и при извлечении информации

Например:

— Напиши запрос в SQL для таблицы [….] сколько заработал [ФИО] за 2020?

— Найти в Уставе коды ОКВЭД

Работа с контекстом

Понимание текста и ответ на вопрос с подсказкой системы – контекстом, индексом – для работы с большим количеством данных

Например:

— Основываясь на данных из […], определи основные тренды развития компании

— На базе […], сформируй список наиболее часто задаваемых вопросов

Какие еще задачи решает GigaChat

Генеративный поиск по базе знаний

Пользовательский интерфейс, где пользователь задаёт вопрос в текстовом виде и получает текстовый ответ и ссылку на документ, в котором этот ответ был найден.

Речевая аналитика

Обработка больших массивов коммуникаций, используя запросы на естественном языке

Автоматизация обслуживания

Облегчение работы по созданию и эксплуатации ботов и диалоговых ассистентов

Определение параметров и поиск аналогов

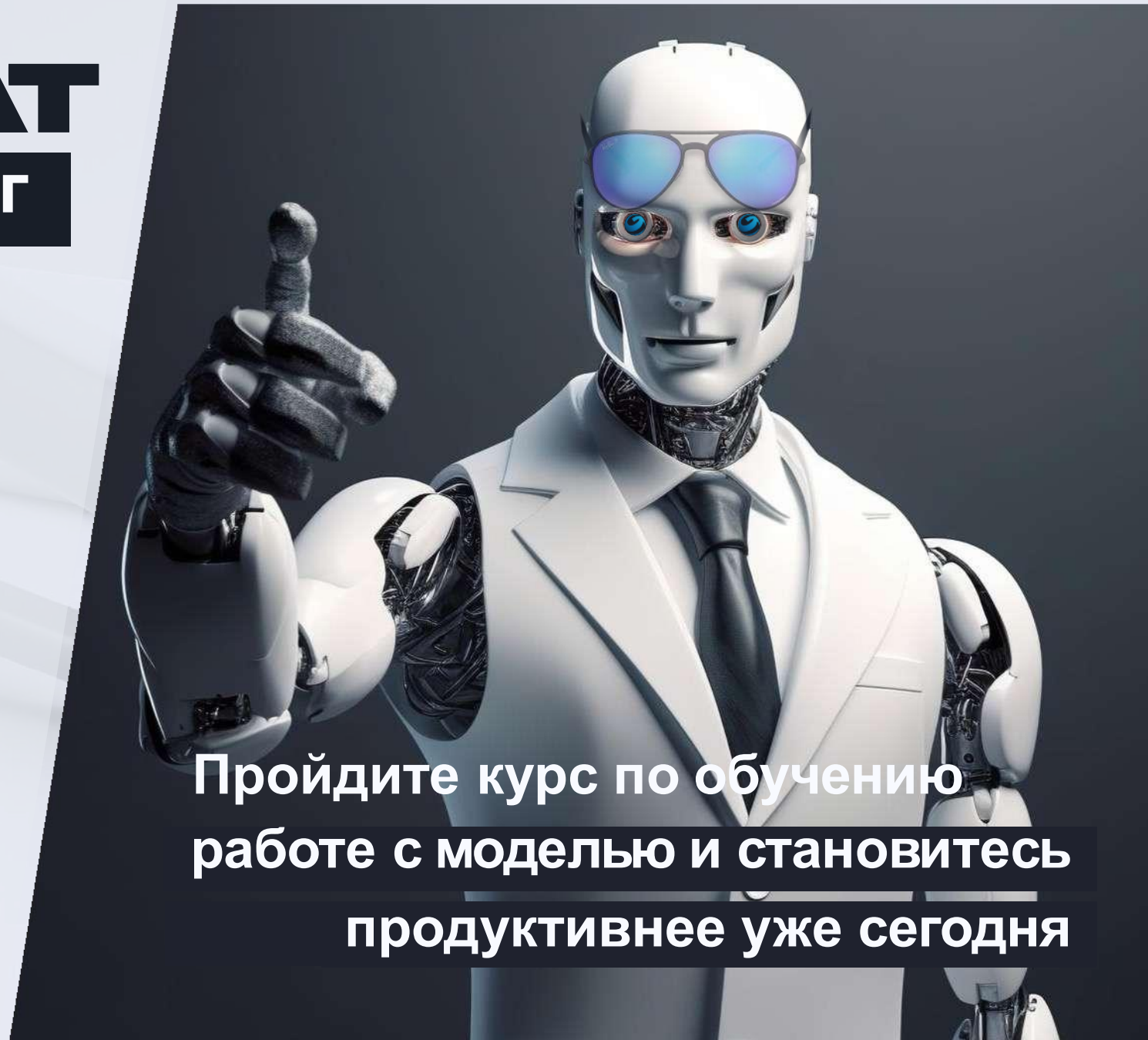
Пользовательский интерфейс, где пользователь указывает текстовое название оборудования и получает карточку с заполненными параметрами номенклатурных позиций из ГОСТ.

GIGACHAT

промпт инжиниринг

Работа с LLM GigaChat

Промт-инжиниринг, сценарии, GigaChain, системные промты, RAG и многое другое ☺



Пройдите курс по обучению работе с моделью и становитесь продуктивнее уже сегодня

Какой вклад генеративных нейросетей возможен?

SBER DEVICES

Создание персонализированного контента

через знания о клиентах и их предпочтения, а также за счет формирования синтетических данных

Формирование подборок и нестандартная классификация

альтернативный поиск по неформальным описаниям (специфика семантического и ассоциативного ряда)

Структурирование информации

быстрый анализ и суммаризация различных объемов данных

Эмерджентность нейросетей

случайно открываемые новые способности модели, которым специально ее не обучали

Исполнение инструкций

обработка обратной связи клиентов, преобразование текстовых данных и их систематизация

Сценарии

КЕЙС 1

Генеративный поиск по базе знаний

SBER DEVICES

поиск по базе знаний

CROSS-INDUSTRY RE-USE

Задача:

Повысить скорость и качества поиска информации по базам данных компании.

Срок реализации: **~16 недель**

Команда: **~18 человек**

Сложность реализации:

Средний

Потенциальный эффект:

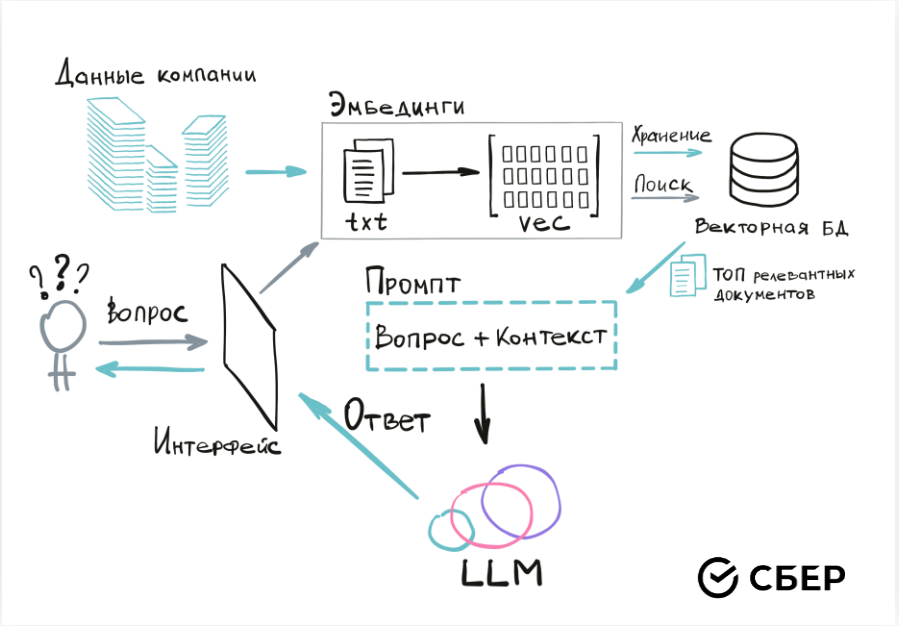
Средний

Решение:

Система умного поиска по базам знаний на основе архитектуры RAG.

Затраты на инфраструктуру:

Отсутствуют



Результат:

Кратное увеличение качества поиска
Точность ответов **90%**

КЕЙС 1

Co-pilot финансового аналитика

SBER DEVICES

Co-pilot финансового аналитика в промышленном предприятии

CROSS-INDUSTRY

Задача:

Сократить время подготовки факторного анализа отклонений МД

Срок реализации: **~18 недель**

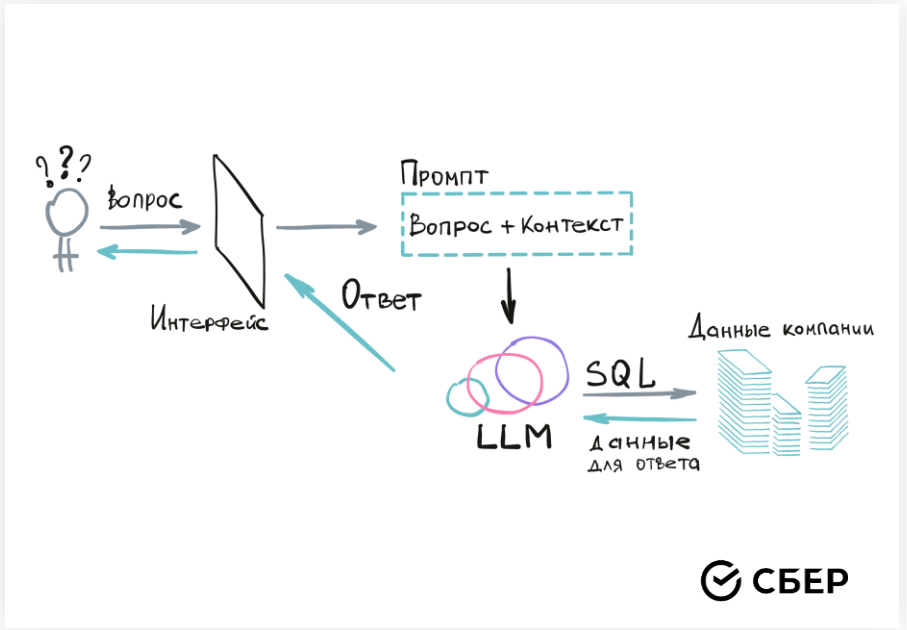
Команда: **~10 человек**

Сложность реализации:

Средний

Потенциальный эффект:

Средний



Решение:

Вопрос-ответная система по базе аналитических данных. GigaChat преобразовывает запрос на естественном языке в SQL-запрос и возвращает точную цифру из базы данных.

Результат:

Кратное увеличение скорости анализа данных
Повышение качества принятия решений
Точность ответов **92%**

Варианты поставки

Широкие возможности модели поставки

Клиенты	МСБ / ISV	Крупный и крупнейший бизнес / Гос. (через Solution-партнеров)			
Тип поставки	Cloud		On-prem		
Продукт	Giga Public	Giga Private	Giga OnPrem	Giga Server	Giga Cube
	Giga Platform	Giga Platform + Аренда железа в облаке	Giga Platform + Evolution/ Giga Stack (на железо компании)	Giga Platform + Evolution/ Giga Stack + Сервер	Giga Platform + Evolution/ Giga Stack + Серверная стойка
Монетизация	Плата за использование	Аренда Private Cloud	Лицензия	Продажа / аренда оборудования	
	RevShare с продажи конечных продуктов на основе GigaPlatform (гипотеза)				

ПАК GIGA CUBE или самостоятельная разработка?

Использование open-source моделей

Лицензионные ограничения и неадаптированный alignment

Нет оптимизации под инференс (увеличение требований к «железу»)

Нет гарантии обратной совместимости при обновлении

Нет возможности добавить в pretrain новые знания

Разработка большой языковой модели с нуля

Собрать данные для pretrain (Петабайты)

Подготовить и разметить инструктивный датасет

Построить суперкомпьютер для обучения модели (множество итераций и тысячи GPU)

Сформировать пайплайн обучения (эксперименты на GPU по несколько месяцев)

10+ млрд ₽

инвестиции

3+ года

время

Преимущества ПАКа GIGA CUBE

Навыки модели адаптированы к потребностям бизнеса

x18

адаптация к русскому языку

x100

экономии

x50

ко времени

Стратегия: «от ускорения отдельных задач к E2E цифровым функциям»

Присоединяйтесь к GigaChat