AI-Run

Платформа ИИ внутри X5

X5 Tech

О нас

В X5 Tech с данными работает более 3 тыс. человек. Накоплена большая экспертиза реализации сервисов с использованием “классического” ML и продуктивизации решений на основе данных. Например, прогнозирование спроса, оптимизация промо, персонализация покупательского опыта и многое многое другое.

О нас

Но почему бы не сделать выйти за рамки традицонных задач ритейла?

Ландшафт инструментов и подходов к работе с данными эволюционирует со страшной скоростью. Кто знал про LLM (Large Language Models) до 2022 года?

Но новые инструменты требуют специфичные знания, специалистов, инфраструктуру. Отсюда берет начало история AI-Run, как платформы.

Почему платформа?

Внутри бизнеса рождаются идеи

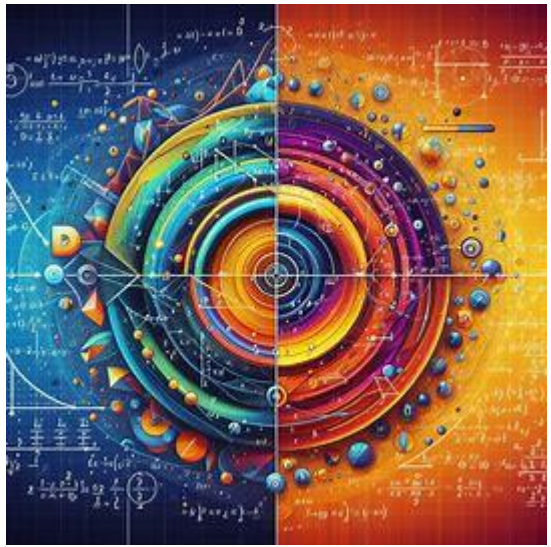
Некоторые очевидные, как поддерживать максимальную свежесть продуктов или не иметь пустых пространств на полках.
Некоторые вполне инновационные

Как минимум часть этих идей мы реализуем с помощью машинного обучения и AI:

1. Классические ML алгоритмы, но на большом объёме данных
2. Рекомендательные системы на нейросетях
3. Модели естественной обработки языка
4. Распознавание документов с помощью компьютерного зрения
5. ...

Как выбрать решение для реализации?

Принятие решения о разработке основывается на разных вариантах внутри и вне организации. Но, даже если стратегическое решение о движении в сторону проекта с ML и AI принято, все еще существуют развилки для решения поставленной задачи.



Если упрощать, то у нас есть два оси координат для принятия решений:

1. Будет ли это универсальное решение, или специализированный подход?
2. Будет ли это внутренняя или внешняя разработка?

Плюсы и минусы коробочных решений

Плюсы

+ **Быстрый старт.** Для решения задачи можно воспользоваться готовой инфраструктурой и набором инструментов;
+ **Масштабирование.** Повторяющиеся запросы можно решать схожими методами. Улучшая одну часть платформы, мы улучшаем показатели сразу нескольких продуктов.
+ **Контроль.** Всегда есть возможность понять, кто что использует. Процесс экспериментов и работы с AI становится прозрачнее.

Минусы

- **Ограниченность возможностей**. Очевидно, удовлетворить всем потребностям одни решением практически невозможно. Не все можно настроить “под себя”;
- **Конфиденциальность данных.** Если в платформе используются инструменты от внешних поставщиков, то всегда есть риск потери чувствительных данных;
- **Зависимость от вендора.** Спустя некоторое время станет довольно сложно отказаться от поставщика решения

Плюсы и минусы внутренних решений

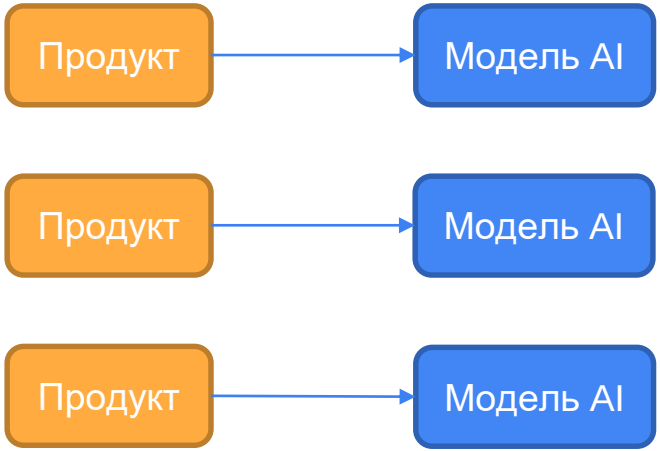
Плюсы

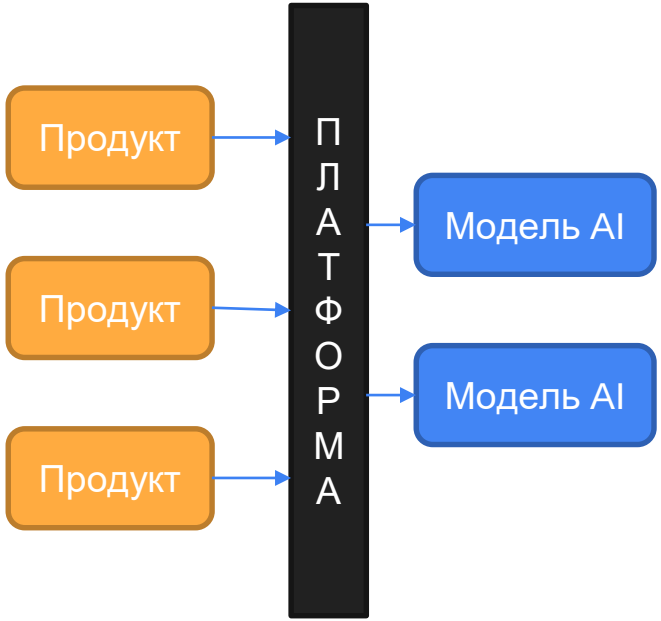
+ **Лучшая подстройка под требования компании.** При разработке можно учесть специфичные требования, которые нужны нашим пользователям;
+ **Накопление компетенций.** Использование внешних решений не позволяет накапливать внутренние компетенции по работе с AI-решениями;
+ **Контроль данных.** В случае внутренней платформы (как и моделей) практически нивелируется возможность утечки чувствительных данных.

Минусы

- **Высокая стоимость начально разработки.** Эффект масштаба работает только внутри компании. Потому затраты на разработку могут быть выше приобретения готового решения;
- **Необходимость специализированной экспертизы.** Обратная сторона накопления компетенций. Чтобы стартовать разработку, нужны достаточно редкие эксперты.

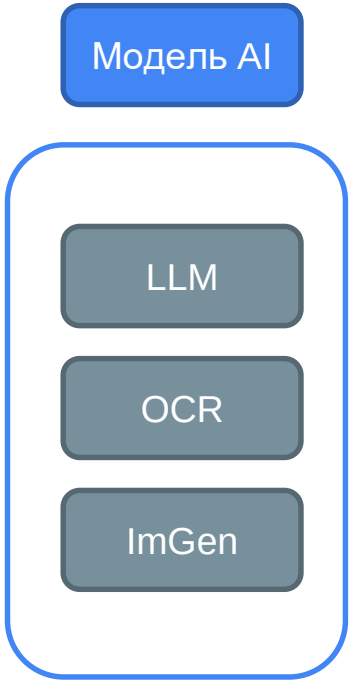
Почему платформа ?





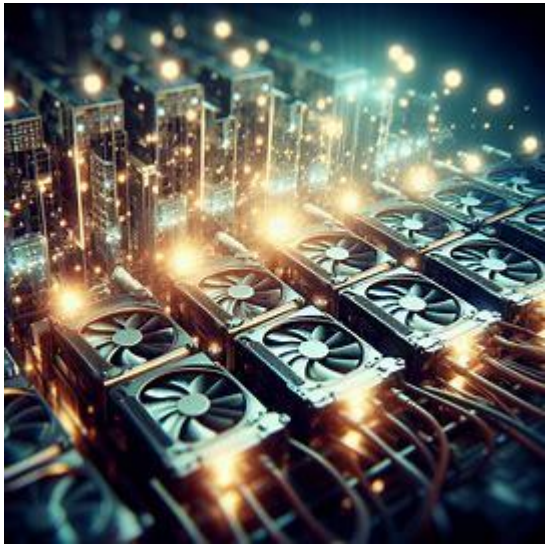
Почему платформа ?

- У команд-пользователей не будет необходимости держать специалистов по AI
- Под капотом можно менять подрядчиков, нет сильной зависимости
- Быстрее можно начать тестирование новых идей
- Долгосрочно на 20-40% дешевле поддерживать каждый отдельный продукт



Почему дешевле и быстрее ?

- Очень часто оборудование закупают с запасом и оно «недоиспользовано». Общий пул оборудования позволит повысить уровень использования
- Не надо будет искать специалистов по AI в каждый проект, это долго
- Согласования и постройка архитектуры будут существенно проще
- Запуск пилотов и проверка идей станут максимально быстрыми



Особенности разработки платформы

Главная особенность – движение “снизу-вверх”. То есть, мы отталкиваемся от потребностей бизнеса. Но, чтобы их получить, нужно провести немалый объем работы.

Что может помочь при таком подходе:

- **Просветительская работа.** Нужно подготовить материалы, которые позволят разобраться в особенностях технологии и сформулировать запрос;
- **Целенаправленный сбор “болей”.** Конечно же, бизнес нужно “подтолкнуть” к генерации гипотез. С этим помогут воркшопы, где целью будет генерация идей о применении AI решений для решения “болей” бизнеса;
- **Песочница.** Немаловажно дать простым пользователям “пощупать” технологию. Это позволит точнее формулировать запросы. Дополнительным плюсом может быть просмотр повторяющихся сценариев применения “песочницы” для перевода их в полноценные сервисы.

Кейсы, на которых строим платформу

Кейсы - QA

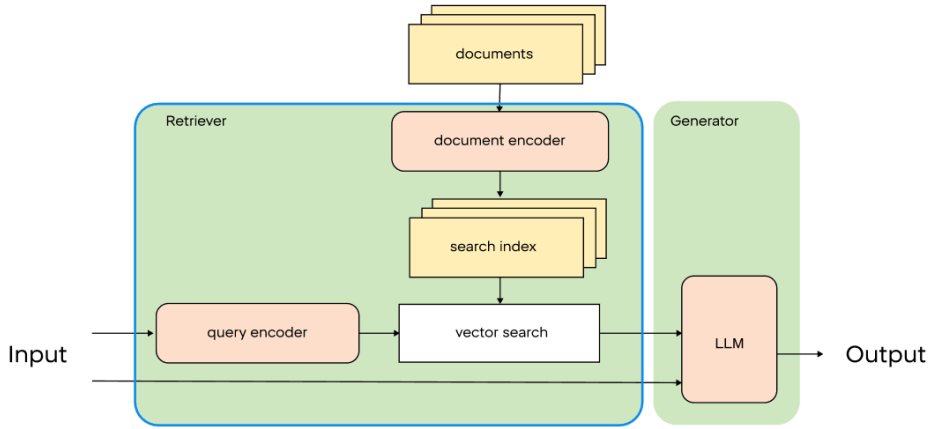
Вопросно-ответная система (QA system) – интеллектуальный инструмент для получения мгновенных и точных ответов по существующей базе знаний. При этом, вопрос может быть задан с помощью “обычной” речи (например, “как списать стул”, “как принять алкоголь” и т.п.).

Учитывая объем накопленных знаний и обширность процессов для каждой бизнес-единицы, такие системы могут сэкономить огромное количество времени.

Как можно решить кейс?

RAG (Retrieval Augmented Generation) – это метод, который в запрос к LLM добавляет “отобранную” заранее информацию. И именно она должна использоваться моделью при ответе.

Обычно, информация для добавления ищется по близости контекста. То есть, мы стараемся найти наиболее релевантную информацию. И потом работать именно с полученными результатами поиска. Чем-то это похоже на то, как с информацией работает человек.



Кейсы – переработка документов

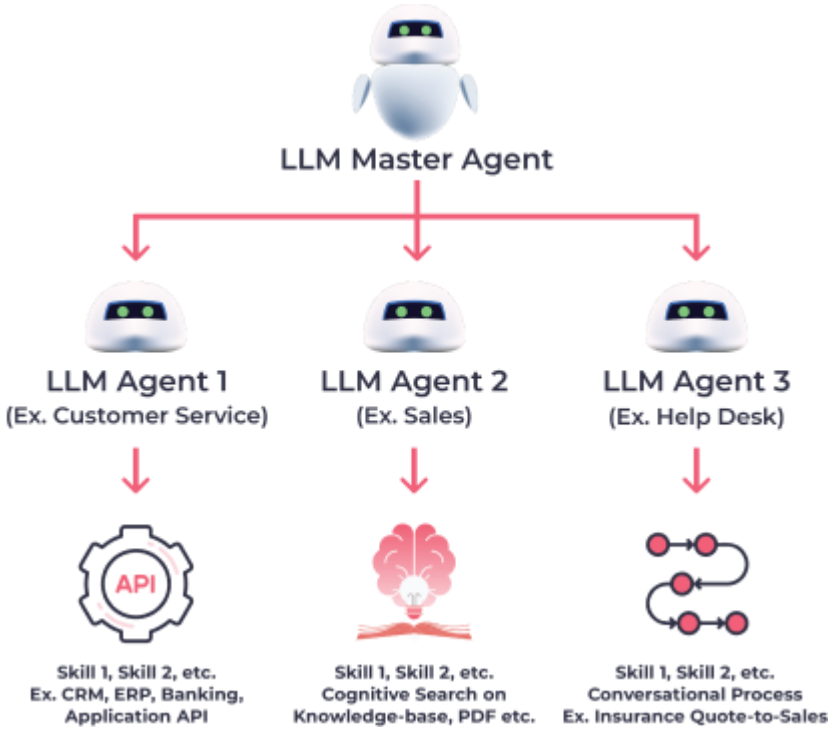
Система помощи в редактировании документов. Еще одна достаточно рутинная задача – это подготовка и редактирование всевозможных документов. Использование LLM позволит:

- Автоматически отслеживать грамматическую и стилистическую правильность текстов;
- Сократить время обработки документов;
- Повысить четкость и понятность текстов;
- Адаптировать тексты к требуемому корпоративному стилю.

Как можно решить кейс?

LLM-агенты – способ использования LLM, при котором мы создаем некоторого виртуального актора, который:

- Отыгрывает определенную роль. Например, можно задать границы “экспертности” нашего агента
- Способен обрабатывать входящую информацию в требуемом виде
- Использовать инструменты, доступные конкретному агенту (например, средства работы с презентациями)
- Реагировать на обратную связь от человека, а также иметь долгосрочную и краткосрочную “память” для поддержания диалога и поддержания действий в рамках роли



Кейсы - суммаризация

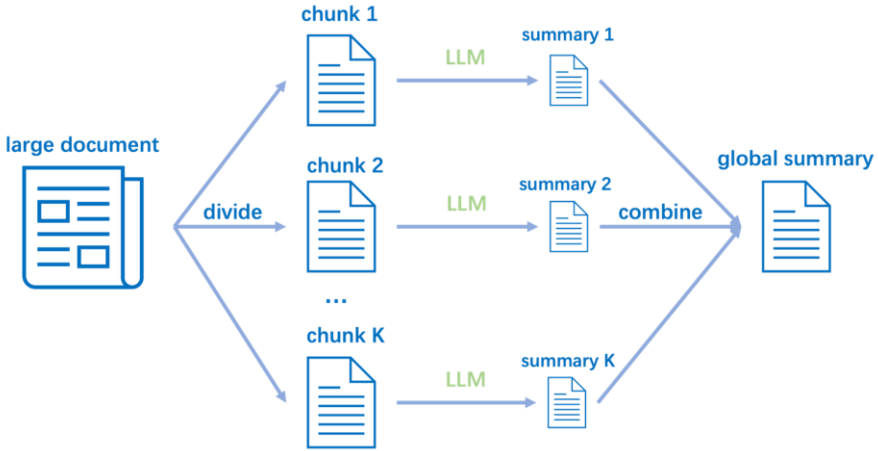
Система выделения сути информации (решение задачи суммаризации). В корпоративной среде достаточно много многостраничных документов, с которыми приходится работать офисным сотрудникам. Логичным кажется выделение сути, что позволяет:

- Качественнее редактировать документы, не теряя суть
- Быстрее подготавливать информационные материалы на основе таких документов
- Проще следить за изменениями смысла (а не формы)

Как решить кейс

Алгоритмы суммаризации развиваются уже достаточно давно. Но LLM могут дать новый импульс таким решениям за счет достаточно больших окон восприятия информации, возможности гибкой настройки промптов для получения саммари и более структурного восприятия текста и сущностей в нем.

Один из типичных вариантов – это тактика “разделяй и властвуй”. Документ делится на меньшие сущности, из которых извлекается краткое содержание. Из получившегося набора кратких текстов комбинируется единый связный текст.



Кейсы - OCR

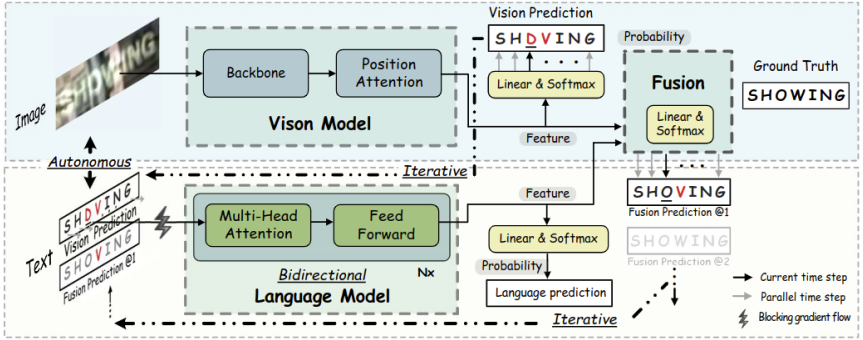
OCR (optical character recognition) – система распознавания текста в разных видах (текстовые файлы, сканы, фото) для последующего перевода в цифровой вид.

Кроме классических задач мы хотим научится распознавать этикетки и составы

Дополнительным плюсом является синергетический эффект при объединении с кейсами на основе использования LLM. Так мы не только извлекаем большой пласт “скрытых данных”, но и сразу же производим интеллектуальную обработку этих данных.

Как решить кейс

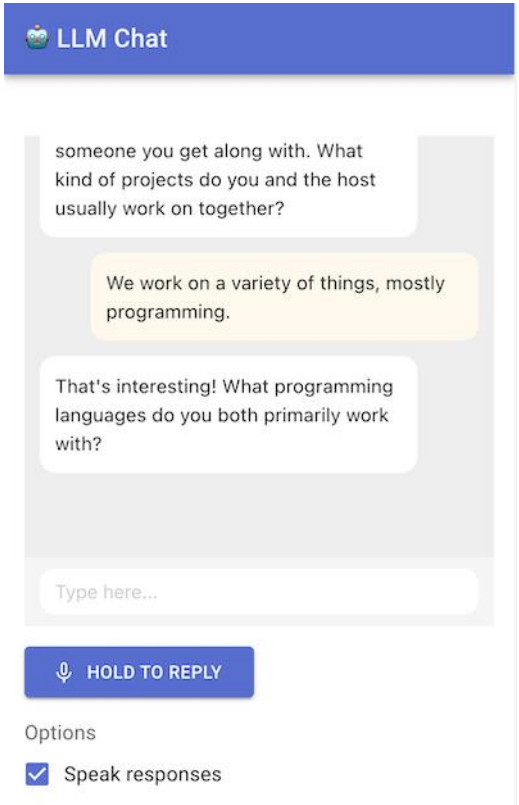
На данный момент задача OCR решена достаточно хорошо. Обычно используются два типа моделей: визуальная и языковая. Одна извлекает информацию из изображения, а вторая корректирует ошибки считывания, учитывая информацию о вероятности получить именно такой текст.
Сложность в создании гибкой системы для разных потребностей



Кейсы – персональное использование

LLM песочница. Позволяет:

- Дать бизнес-пользователям “потрогать” технологию
- Снизить нагрузку на команду (пользователь может самостоятельно решить часть своих запросов)
- Выделить повторяющиеся запросы для последующей автоматизации и превращения в сервис



Итоги

Итоги

Подведем итоги сегодняшнего рассказа:

- Даже при наличии большой экспертизы, работать с наиболее современными достижениями в сфере AI может быть непросто
- Наше решение для задачи получения возможности использования таких технологий – AI-платформа
- При этом, реализация такого рода платформы имеет свои плюсы и минусы. Каждой организации стоит выбирать свой способ реализации потребности в работе с достижениями в сфере AI