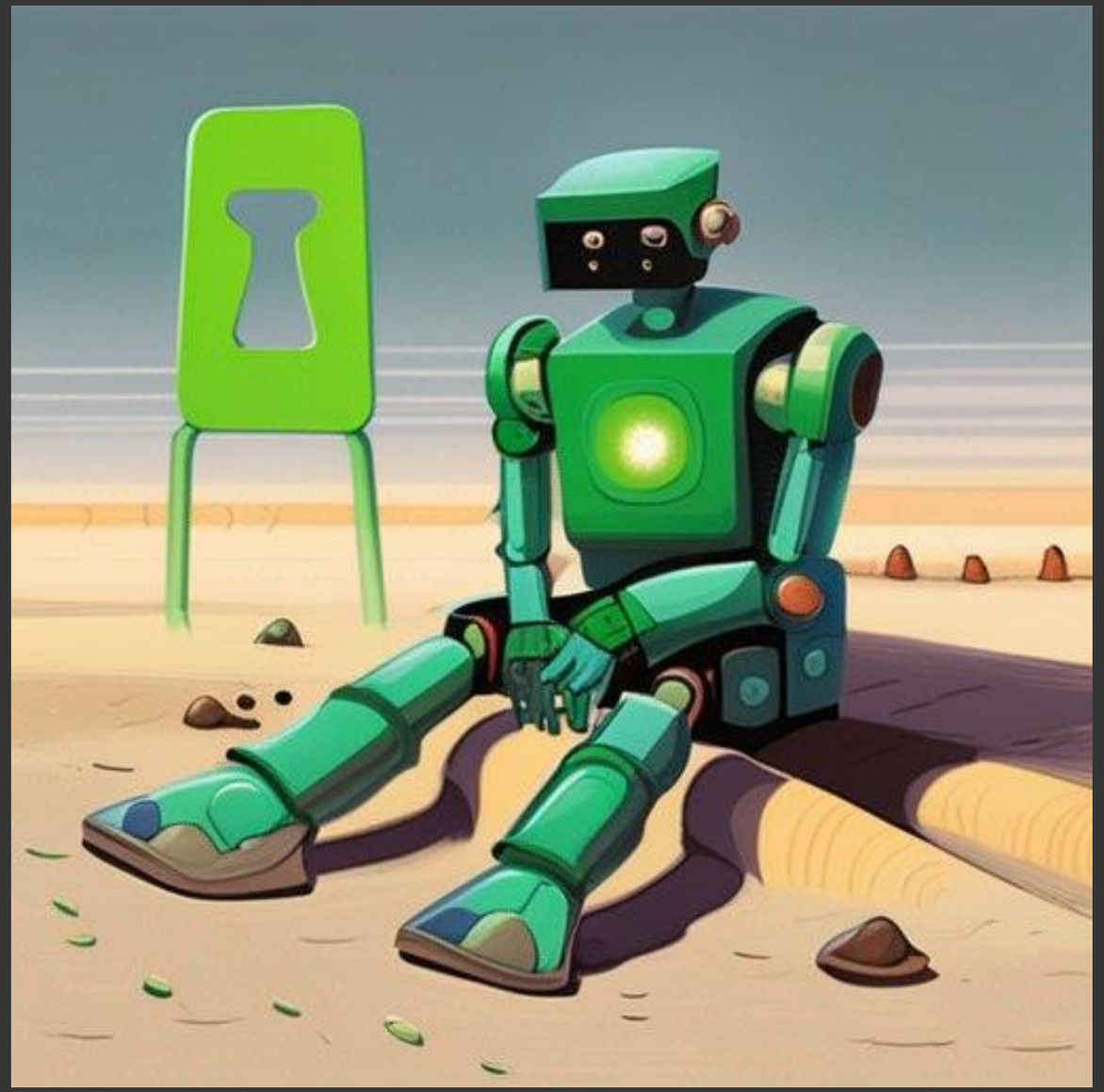
Чат-боты
на основе Retrieval-Augmented Generation



Дмитрий Сошников
Доцент, НИУ ВШЭ/МАИ
Консультант Яндекс по ИИ и машинному обучению
http://soshnikov.com – @shwars
http://t.me/shwarsico

О чем будем говорить

YandexGPT

DataSphere

LangChain

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[⋆],

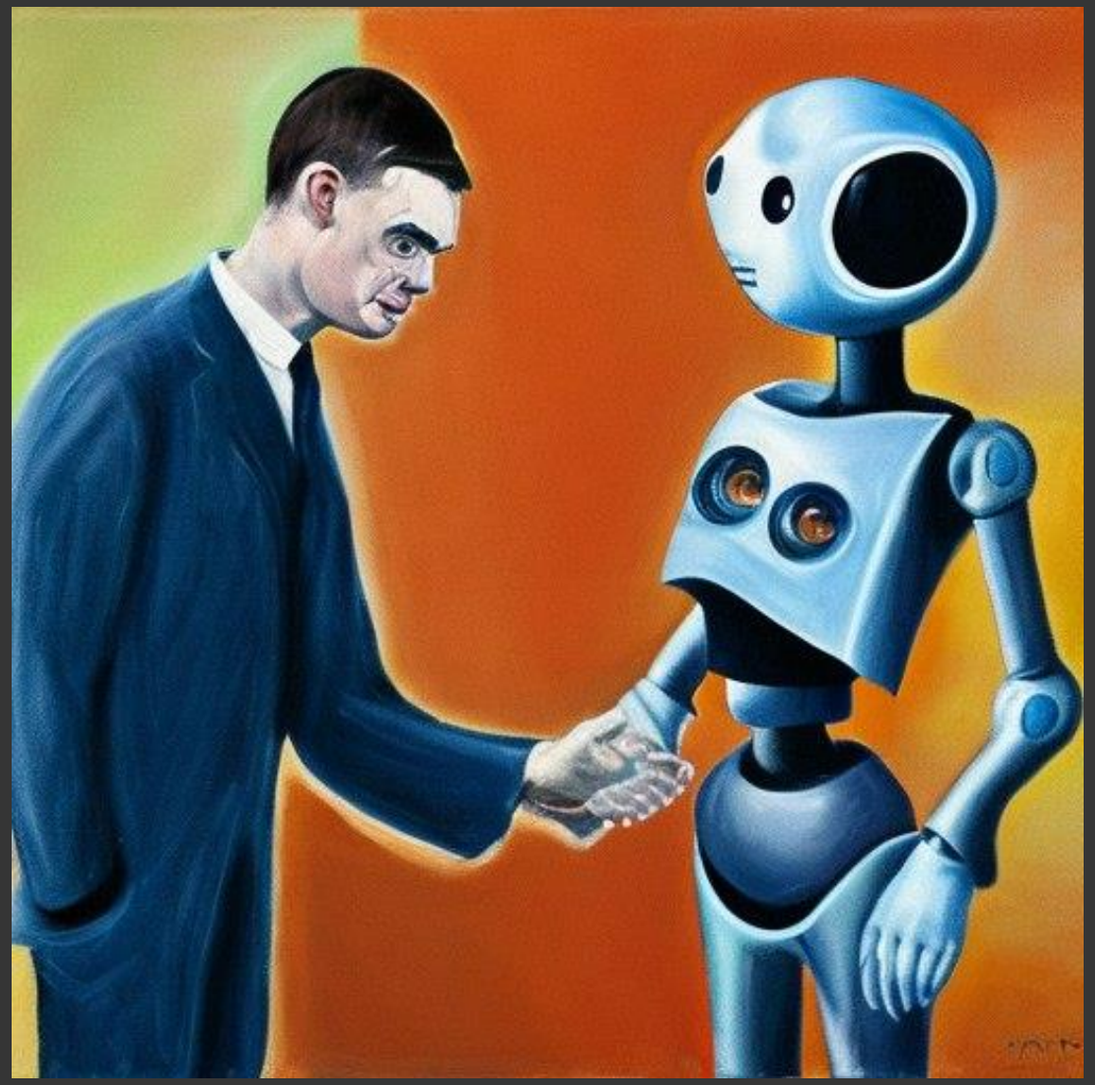
Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

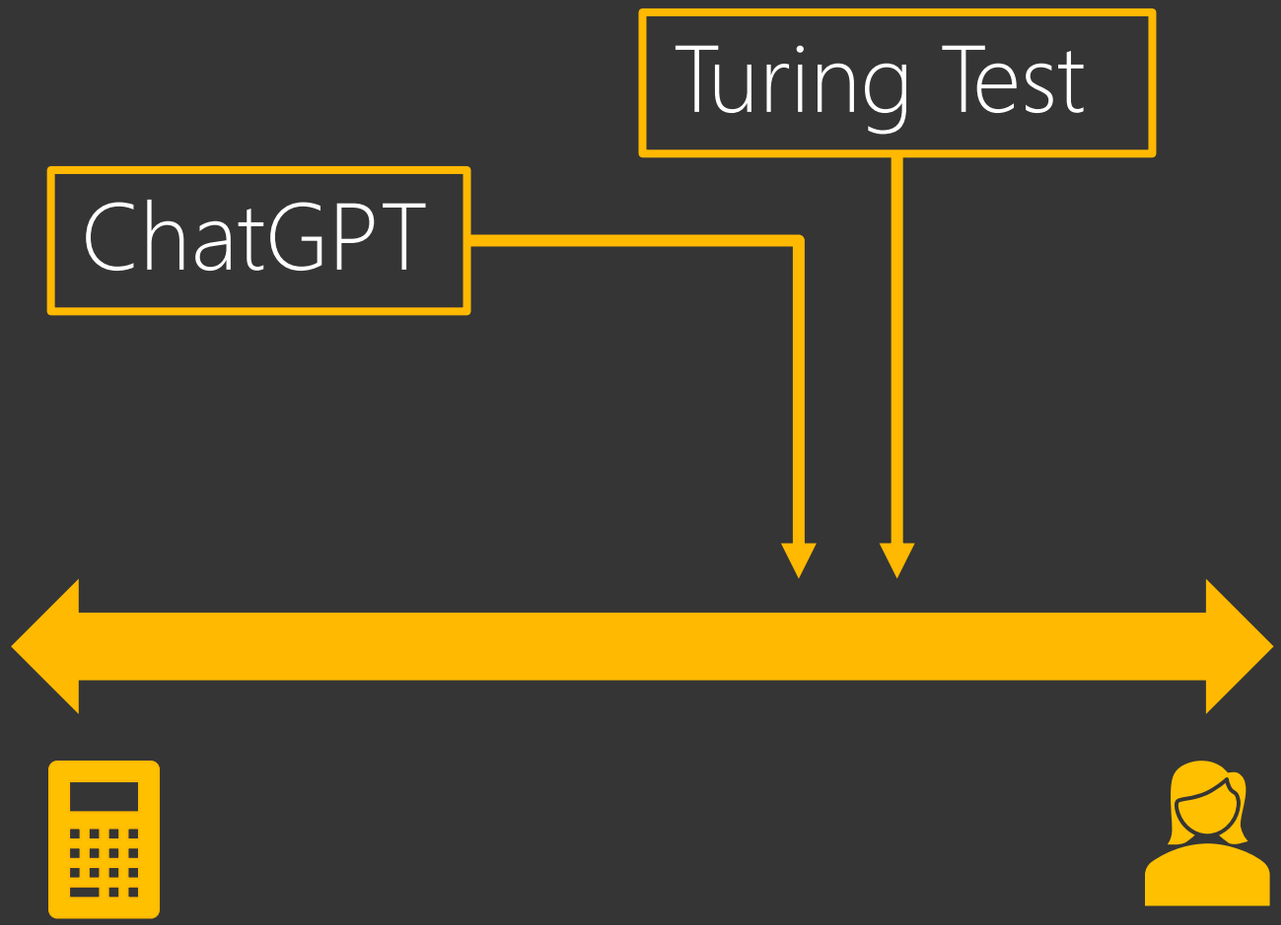
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;
plewis@fb.com

Самое главное: насладиться GPT+DataSphere

Искусственный интеллект





GPT

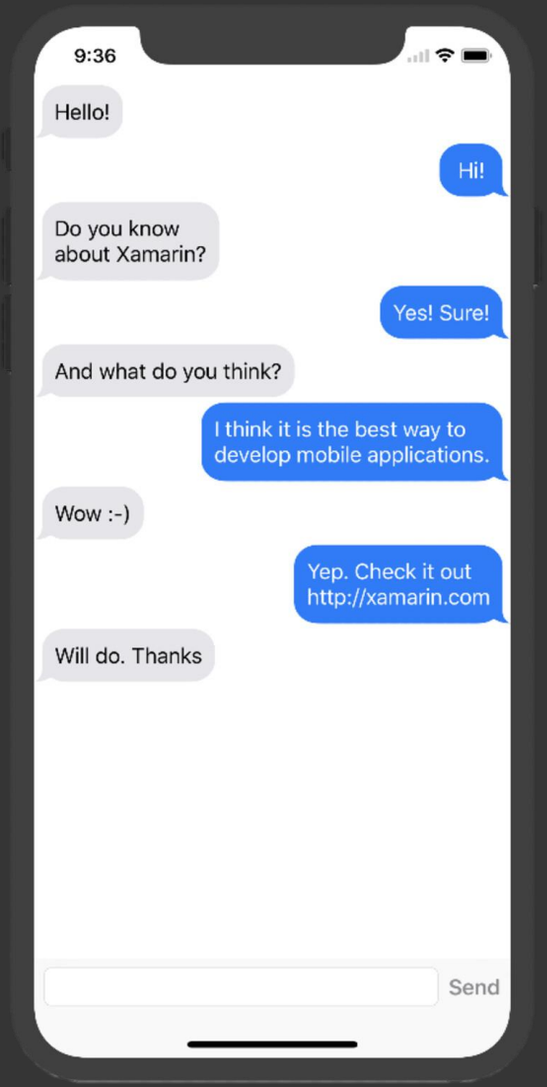
Британские исследователи университета им. Джона Леннона в Ливерпуле на прошлой неделе закончили исследования, посвященные влиянию вируса на музыкальные предпочтения английской молодёжи. Они обнаружили, что у людей, заражённых вирусом иммунодефицита человека, музыкальные вкусы изменились. В результате, по мнению специалистов, в будущем, как минимум, половина британцев будет слушать музыку с преобладанием рок-мелодий. В исследовании принимали участие более 100 молодых британцев в возрасте от 14 до 18 лет. В результате оказалось, что среди участников эксперимента, заражённых вирусом иммунодефицита, рок-музыку слушали только 13%, а в группе, заражённой не-ВИЧ-вирусом - только 7%. По словам профессора Гарварда Майкла О'Салливана, результаты исследования говорят о том, что музыкальное развитие молодых людей в Великобритании находится в прямой зависимости от наличия у них вируса иммунодефицита человека. «Наше исследование показывает, что если у людей, заражённых вирусом иммунодефицита человека, музыка будет преобладать в жизни, то у тех, у кого его нет, она не будет иметь особого значения. И наоборот, если у молодых людей будет преобладать рок-музыка, то в будущем их музыкальные вкусы, вероятно, изменятся», - сказал профессор.

Цепи Маркова

Мой дядя самых честных правил...

Проект объёмный, над ним работает несколько технологов. Чтобы посетители не вытоптали остатки леса, за заграждения заходить нельзя. Они как бы ограничивают распространение изменений. Мы пишем такие «переходники», которые делают опасное вождение *неудобным*.

Экспертные системы



Yandex GPT как экспертная система

Какой макбук лучше выбрать студенту?

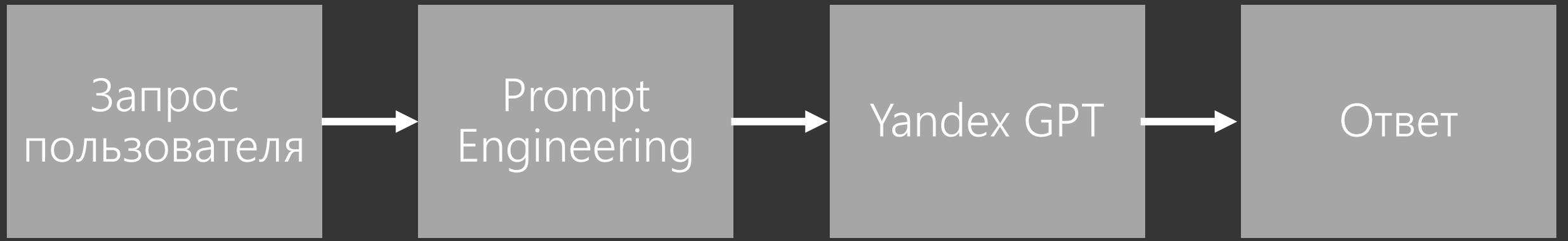
Выбор Макбука для студента зависит от нескольких факторов, таких как бюджет, требования к производительности и задачам, которые предстоит решать в университете. Важно учитывать технические характеристики, такие как процессор, оперативная память, жесткий диск, а также другие особенности, такие как размер экрана, графика и другие. Если студенты ищут недорогое устройство, то Apple Macbook Air может быть хорошим выбором. Он имеет легкий корпус, маленький экран и низкое электропотребление, что позволяет использовать его в дороге с легкостью. Однако, если нужен более мощный ноутбук для рабочих задач, тогда стоит рассмотреть Macbook Pro. У него более производительный процессор и большой объем оперативной памяти, что позволит выполнять сложные задачи без задержек. Также стоит учитывать, что экран у Macbook Pro больше и разрешение выше, что делает работу более комфортной. Наконец, если необходимо иметь ноутбук с возможностью расширения оперативной памяти и хранения данных, следует обратить внимание на Macbook Pro с дополнительным слотом для расширения. Это обеспечивает дополнительные возможности для производительности при работе с требовательными приложениями.

Минутка рекламы

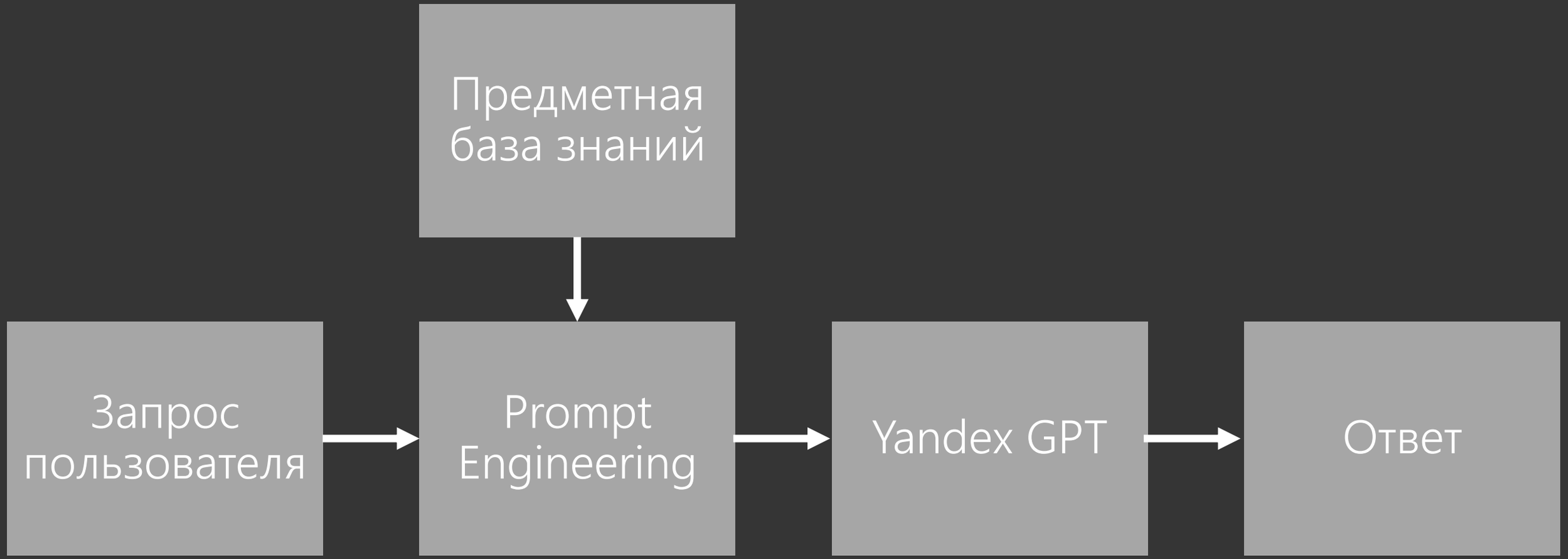
17:30, зал 1

GPT как персональный раб разработчика

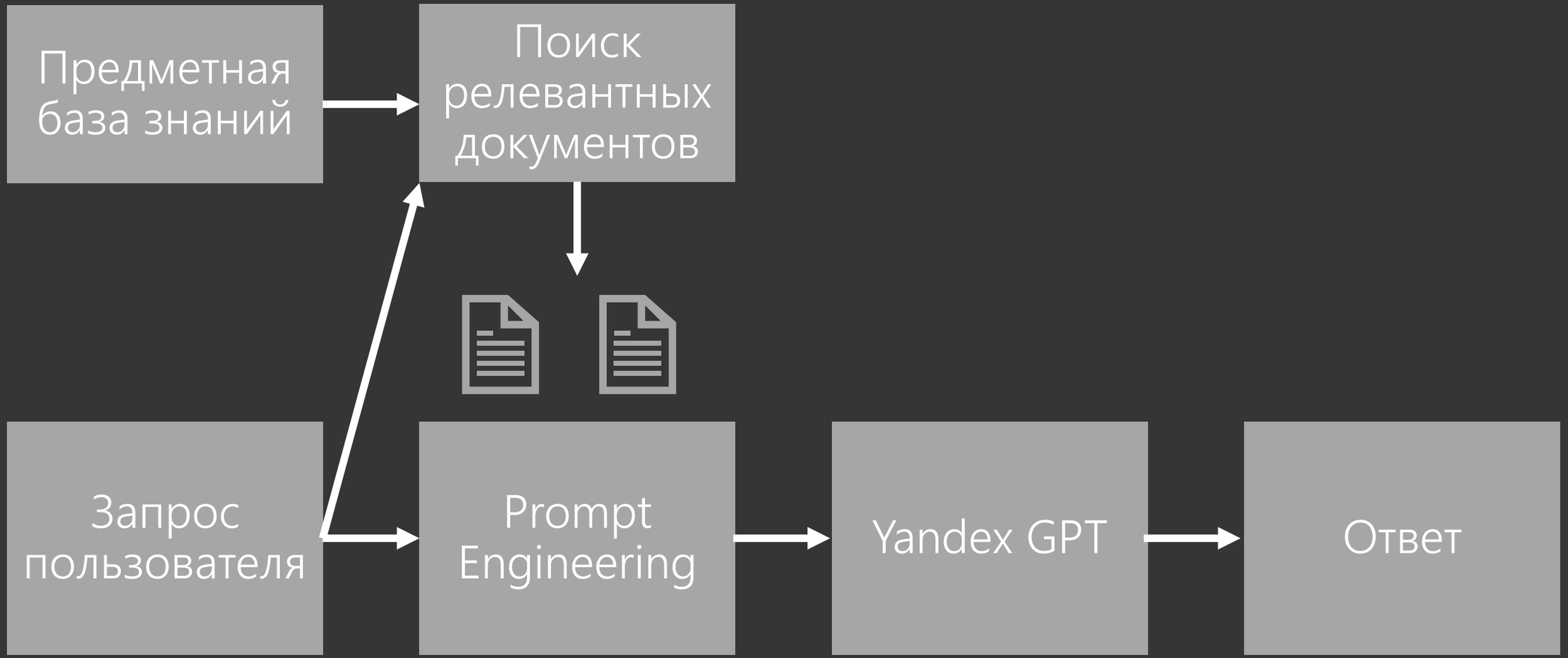
Вопрос-ответный чат-бот на базе GPT



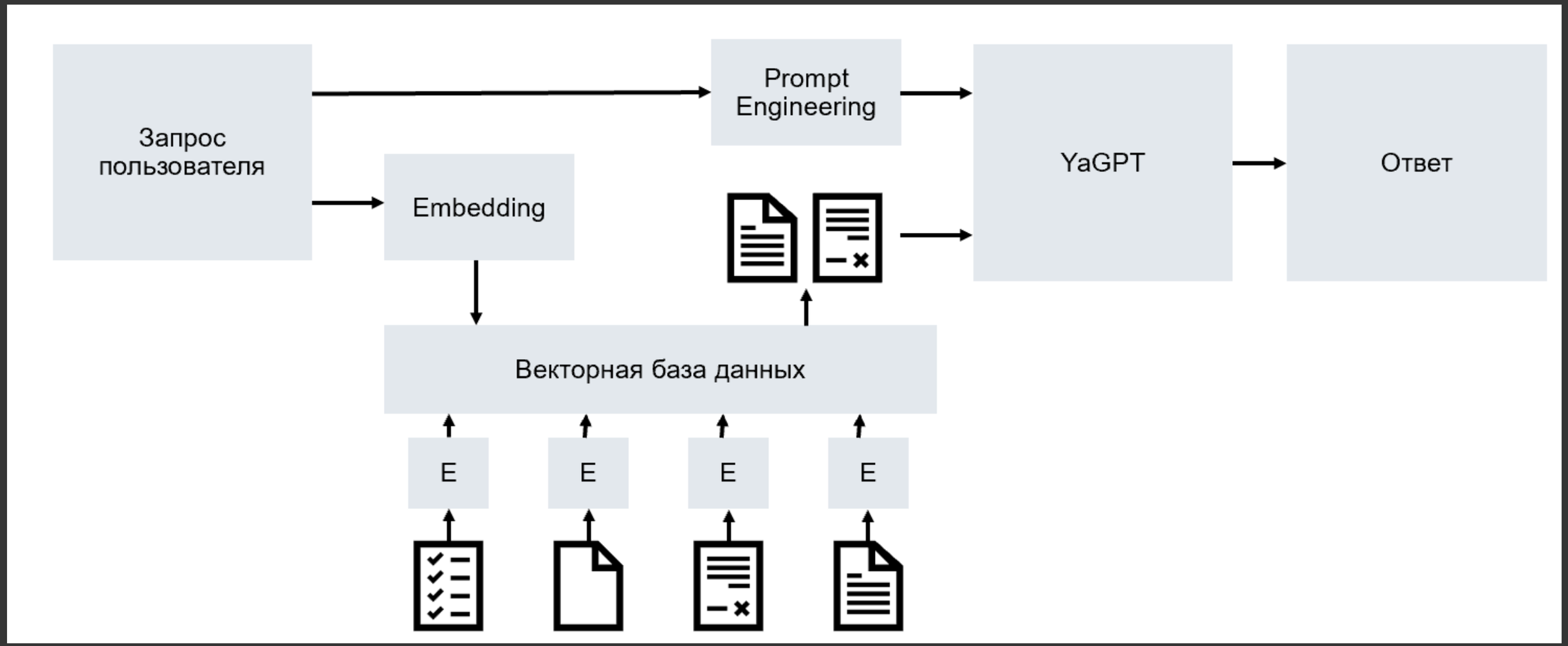
Content-Augmented Q&A



Retrieval-Augmented Generation



Retrieval-Augmented Generation



Альтернативный подход

Fine-tuning

- Дорого и долго
- Необходимы большие объемы текста с разнообразными запросами
- Для внесения изменений нужно повторное обучение
- Отличные результаты, если всё сделано правильно

Retrieval-Augmented Generation

- Быстро и дёшево
- Текст, содержащий нужную информацию
- Для внесения изменений достаточно поменять текст
- Результаты не всегда идеальны

Возможны галлюцинации

Пример: Video Q&A Bot

Yandex Speech Kit

Demo Time!

ML в Yandex Cloud

Готовые модели: доступ по API

SpeechKit

Распознавание и синтез речи. Создание голосовых помощников, автоматизация колл-центров, контроль качества сервиса

YandexGPT

Генеративная языковая модель для анализа текстовой информации, создания контента и чат-ботов

Translate

Машинный перевод на 90+ языков, собственный глоссарий

Vision

Распознавание текста и шаблонов документов, классификация изображений, распознавание автомобильных номеров

Среда разработки и обучения ML моделей

DataSphere

Функционал для командной работы и ролевая модель

JupyterLab

Гибкая конфигурация вычислительных ресурсов с GPU V100 и A100

Data Proc

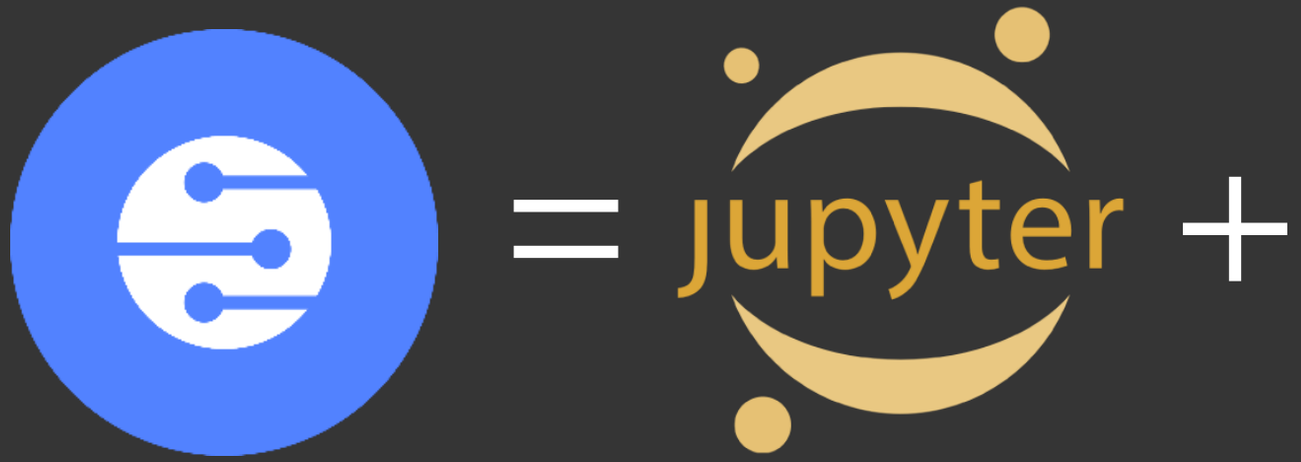
Тесная интеграция со Spark для обработки больших данных в S3

Доступ в Yandex Cloud

https://t.me/dsoshnikov

Yandex DataSphere

https://datasphere.yandex.ru



- Гибкие вычислительные ресурсы
- Датасеты и подключение к S3
- Групповая работа и разделение ресурсов
- Git-интеграция
- DataProc для работы с большими данными

Yandex Speech Kit

Реал-тайм взаимодействие

Голосовые роботы, умные ассистенты, подсказка оператору

Озвучка контента

Аудио книги, новости, подкасты, статьи на сайте, озвучка видео, адаптация сайта для людей с нарушением зрения

Распознавание файлов

Записи встреч и конференций, генерация субтитров, контроль качества

Распознавание коротких файлов

Аудио сообщения в мессенджерах

LangChain

Поддержка в одном флаконе:

- Генеративные языковые модели (llms)
- Вопрос-ответные модели (chat models)
- Эмбеддинги
- Векторные базы данных
- Работа с документами, обработка текста
- Цепочки (chains)
- Агенты
- …

http://soshnikov.com

@shwars