5 остановок

до LLM в твоем проде

Викулин Всеволод I Яндекс

Руководитель службы быстрых ответов

Зачем нам LLM?

С чего начать?

Как оценить качество?

Как выбрать архитектуру?

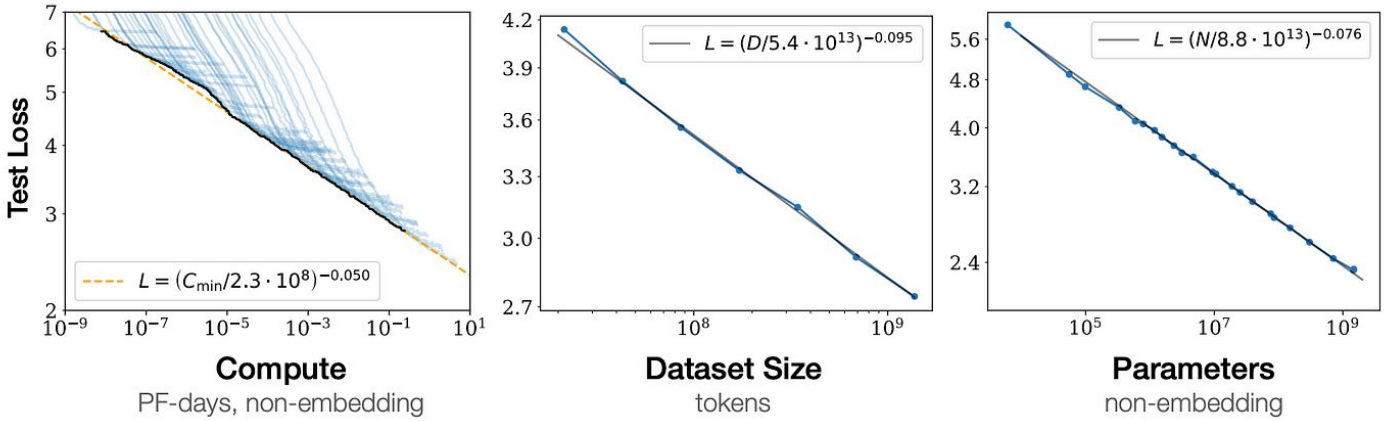
Как сэкономить?






LLM BUS


Остановка 1. Вам точно нужна именно **L**LM?



Scaling Laws for Neural Language Models **(Kaplan et al., 2020)**

Обобщаемость = решает широкие задачи + быстро понимает новые

Чем больше нужна обобщаемость, тем больше L

Разработка кода
Deep Research

30B

Вопрос-ответная система
Простой чат бот

3B

Классификация
Выделение сущностей

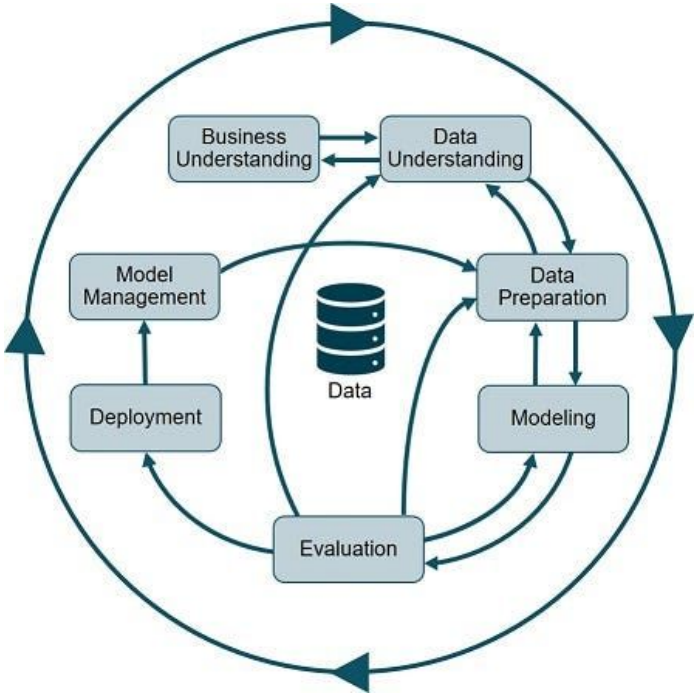
300M

А зачем GPT4 1.5T?

Даже у Альтмана нет бесплатных завтраков

Здесь подробная дока про выбор модели

Остановка 2. С чего начать LLM проект?



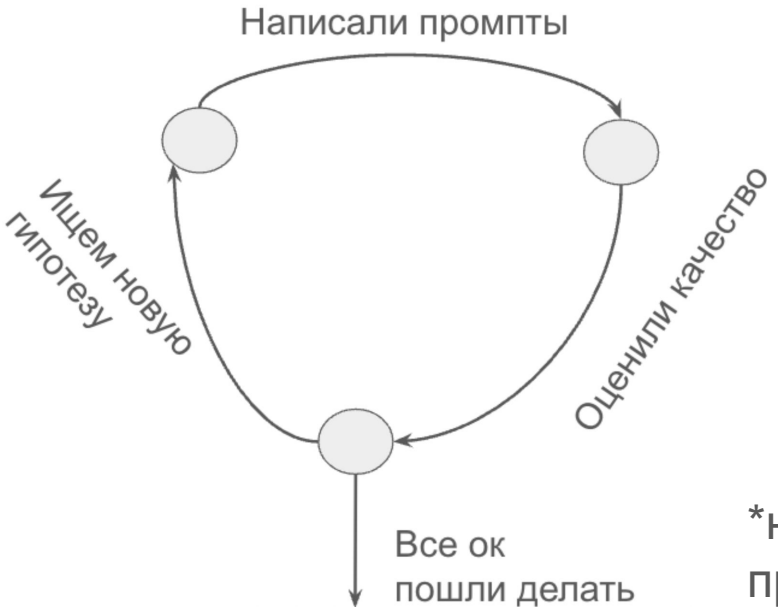
CRISP-DM
Cross-industry standard process for data mining

1. Понимаем задачу бизнеса
2. Собираем кучу данных
3. Учим модель
4. Понимаем, насколько бизнесу похорошело
5. Не очень, потому что мало данных/криво обучили модель, идем в пункт 2
6. …

   …

100. На самом деле и идея была так себе

Prompt driven development



1. Понимаем задачу бизнеса
2. Промптим* прототип за несколько недель
3. Понимаем, насколько бизнесу похорошело
4. Если не очень, ищем ~~другой бизнес~~ другую задачу
5. Иначе начинаем упрощать/ускорять

*не пытайтесь промптить без продакт менеджеров

ПДД
Prompt driven development
Разрешается использовать модели безумного размера
Лучше с ризонингом

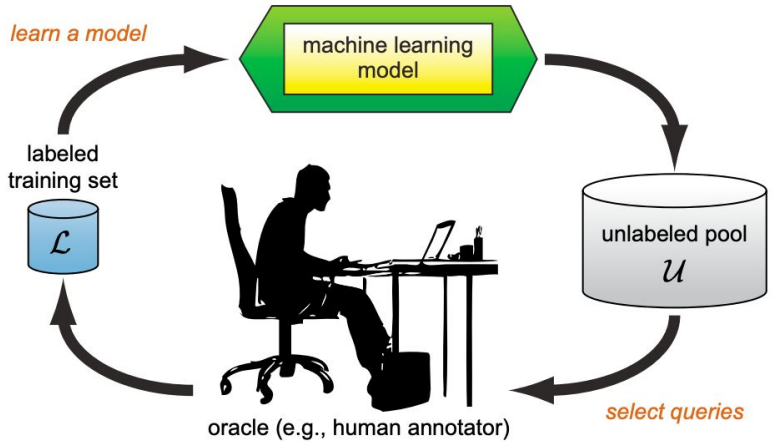
Остановка 3. Как оценивать качество?

Задачи бывают:

1) С правильным ответов (классификация, перевод). Тут легко.
2) Без правильного ответа (почти все остальное). Тут **очень больно**

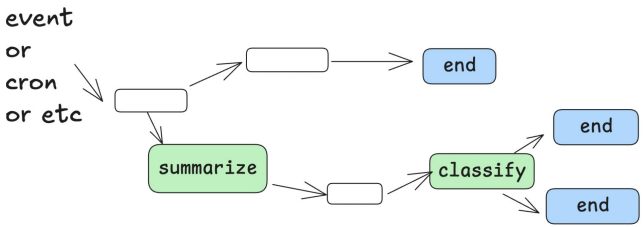
Как учить модели качества

1. Начинаем с LLM-промпта, скорее всего быстро упретесь
2. Аккуратная инструкция, умные разметчики, высокий контроль
3. Дообучаете LLM-классификаторы
4. Используете **active-learning** для сбора примеров



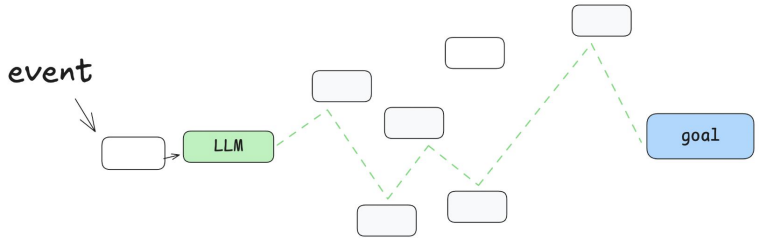
Остановка 4. Архитектура LLM проекта

Старая архитектура (ml workflow)



Написал все ифаки,
ничего не ломается, сижу
довольный

Стильно-модно-молодежно (агенты)



Не писал ифаки, иногда
ломается, сижу
довольный

Тут написал как
взять лучшее из
двух миров

Когда **реально** нужны агенты

Как понять, что пора

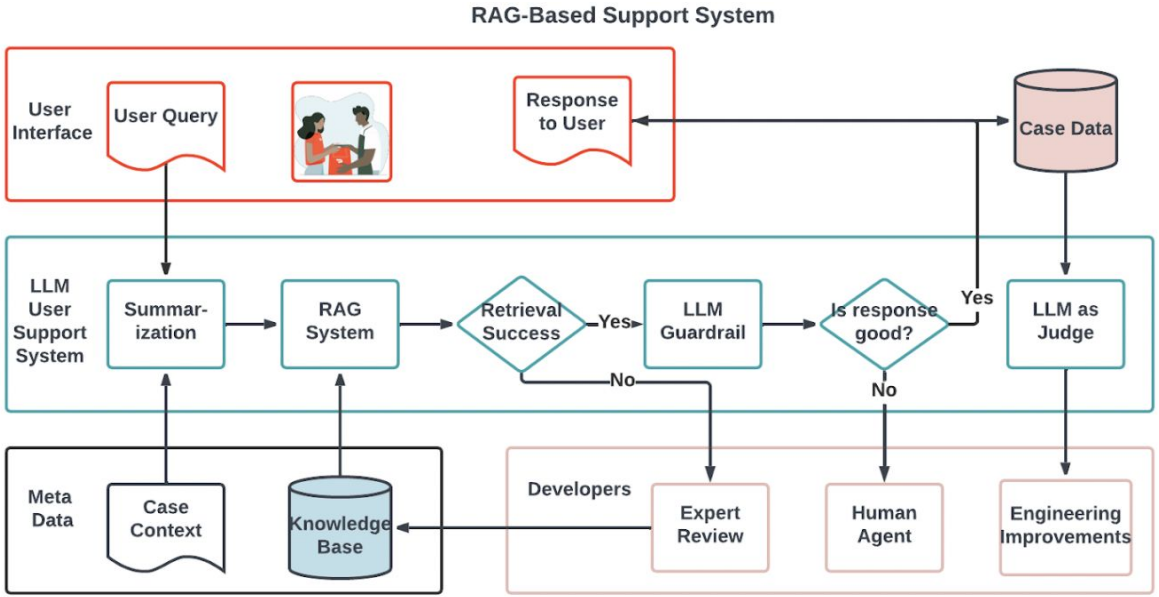
- Я не смогу за адекватное время написать столько логики
- Если оно ошибется, мне ничего не оторвут
- Человек может немного подождать, пока оно подумает



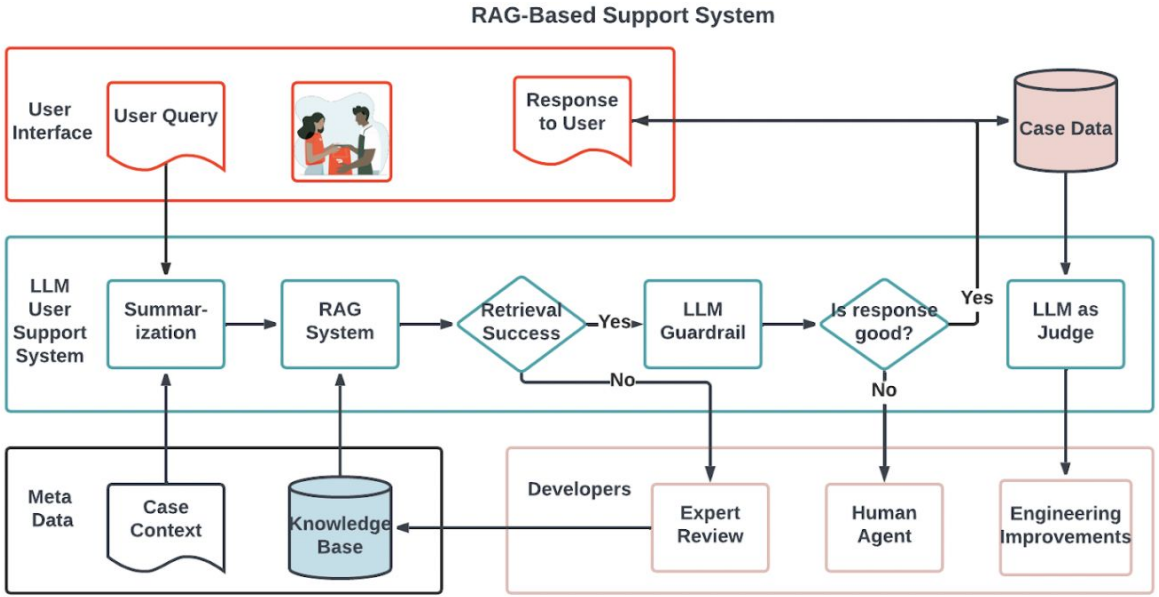
Примеры областей

- **Разработка** - все баги ловим тестами
- **Deep Research** - пользователь сам вычитает
- **Тыкание мышкой** - все оплаты попросим подтверждать, изолированное окружение

Пример реального LLM проекта - DoorDash

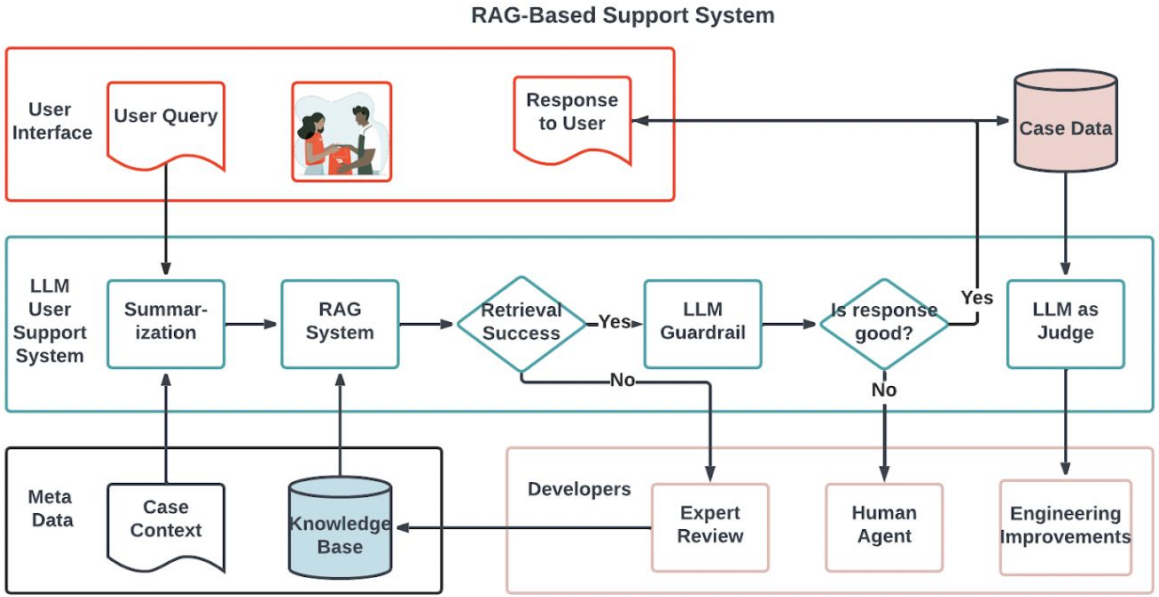


Пример реального LLM проекта - DoorDash



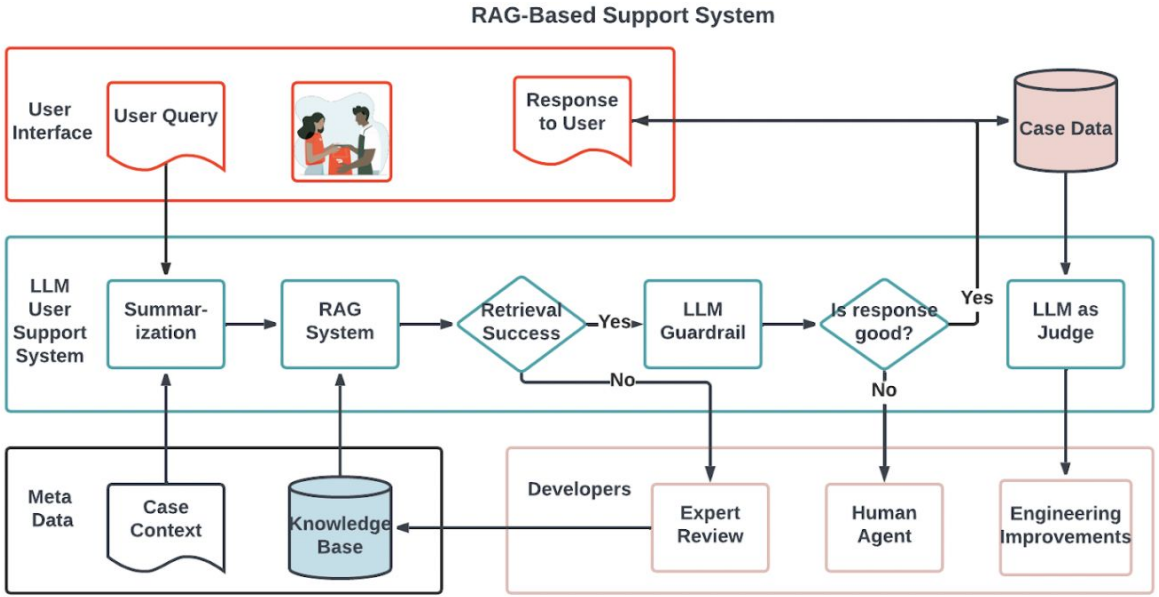
Мысль 1. Всегда дробите на атомарные части

Пример реального LLM проекта - DoorDash



Мысль 2. Всегда используйте incontext learning

Пример реального LLM проекта - DoorDash



Мысль 3. Используйте Guardrail (реворд) на ответ и human-in-the-loop

Остановка 5. Как сэкономить

Шаг 1. Дистилляция.

В вашем проекте самое важное **запустить мотор.**

И остальные наши друзья

1. Квантизация - **How is LLaMa.cpp possible**
2. Continuous-батчинг - как работает
3. Спекулятивный декодинг - оригинальная статья
4. Оптимизация аттеншиона
   a. Flash attention - туториал с имплементаций
   b. Paged attention - объяснение самих авторов
5. Кеширование
   a. Kv-cache
   b. Context caching

Что читать про LLM system design

- Чуть больше про стратегию внедрения LLM
- 500 кейсов внедрения LLM
- Как делать LLM приложения от Anthropic
- 12 принципов проектирования LLM агентов
- Агенты, пожалуйста, понадежнее
- show me the prompt
- Гайд как улучшить AI продукты

СПАСИБО