X5 Digital

Интерактивные рекомендательные системы на LLM

Карина Садова, Head of AI @ X5 Digital

X5 Digital

Про что будем говорить сегодня?

1

Про сравнение классических и интерактивных рекомендательных систем

2

Про то, почему мы не отказываемся от одних в пользу других, а дополняем

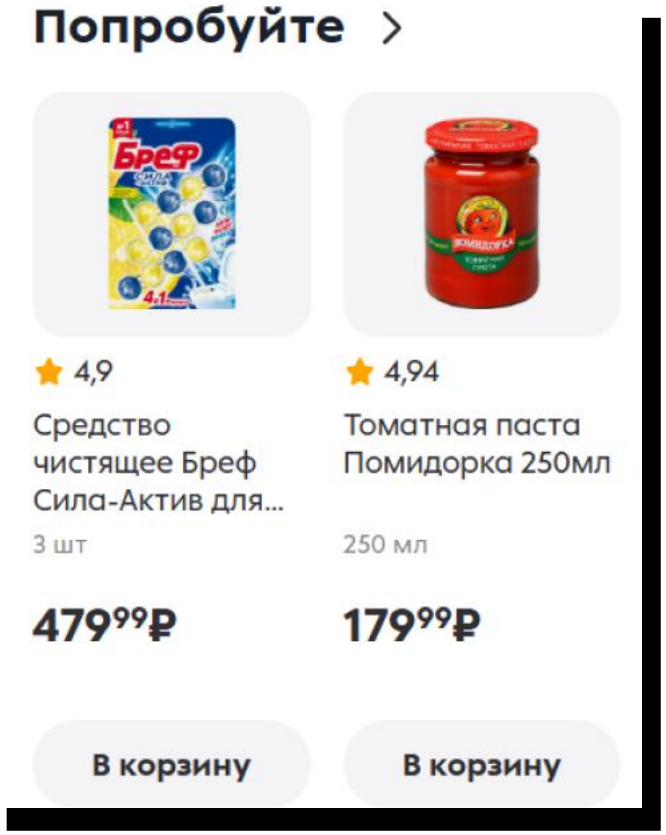
3

Про то, как выстроить бейзлайн этого дополнения на LLM-ках

X5 Digital

Что такое интерактивная рекомендательная система

Классическая рек. система

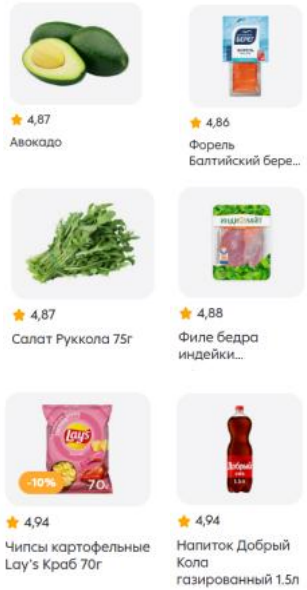


Интерактивная рек. система

“Хочу красиво позавтракать”

“Я теперь на диете, собери мне рацион на 1500 ккал”

“Ко мне сегодня придут друзья на вечеринку, давай накроем стол”



X5 Digital

Кейсы использования интерактивных рек. систем

мы хотим в явном виде сообщить системе о своих потребностях

некоторое редкое, неповторяющееся событие

новое условие, которое будет повторяться

новое условие, которое будет длиться ограниченный период времени

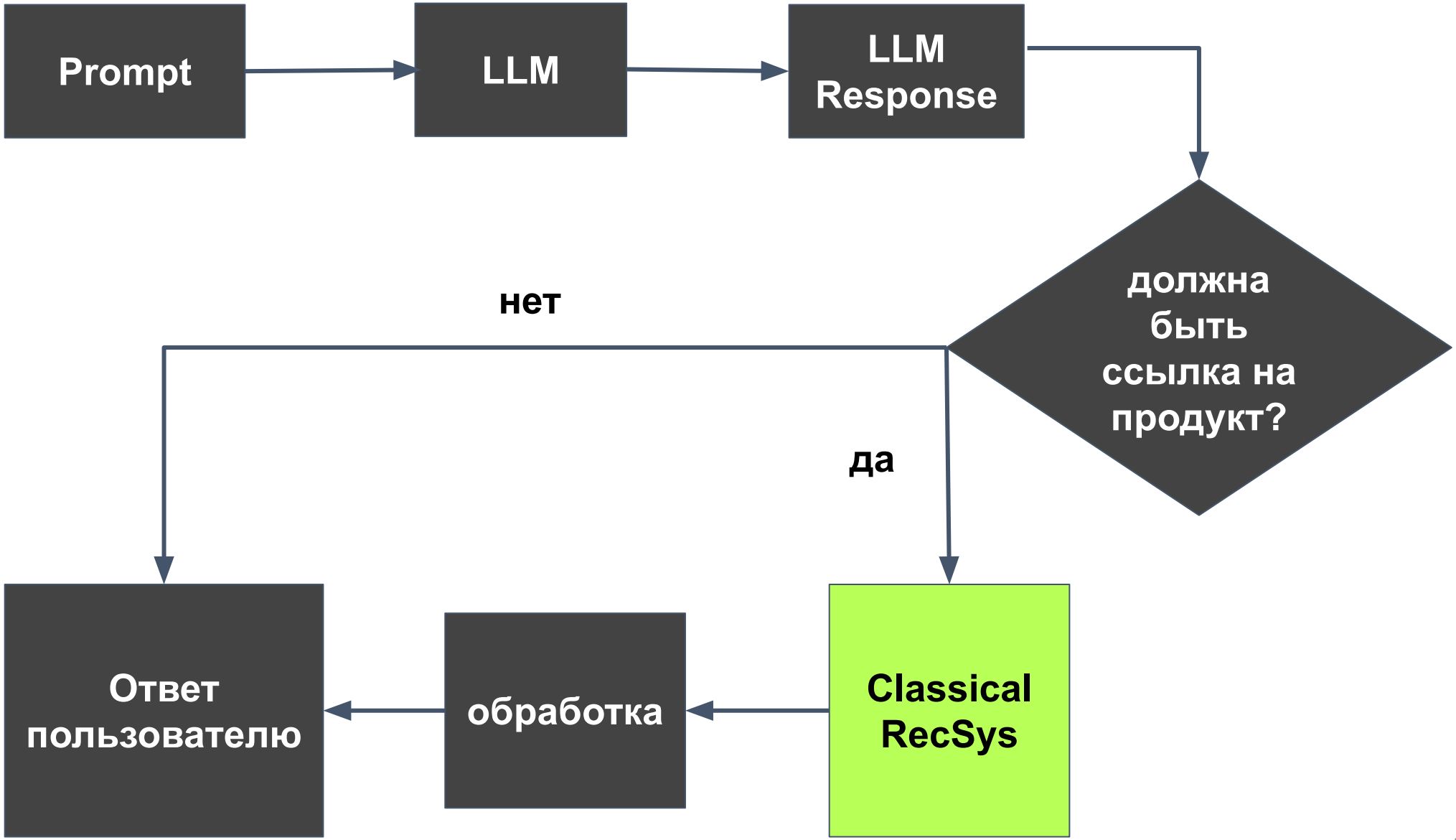
событие, зависящее от настроения или других переменчивых факторов

X5 Digital

Почему мы не отказываемся от классических рекомендаций?

- классические рекомендации хорошо справляются с выбором конкретного товара для конкретного пользователя
    - хлебушек X из всех видов хлеба для пользователя Y
- языковые модели хорошо справляются с интерактивной составляющей
- для компании круто, когда внутри одних продуктов нет велосипедов других продуктов

X5 Digital


Prompt
LLM
LLM Response
должна быть ссылка на продукт?
нет
да
Ответ пользователю
обработка
Classical RecSys

X5 Digital

LLM ответ пользователю

{

Судя по тем продуктам, которые я распознал в твоей корзине, ты можешь приготовить лазанью, греческий салат и глинтвейн.

Но тебе не хватает некоторых позиций. Добавить их?

}

X5 Digital

LLM ответ пользователю

{

Судя по тем продуктам, которые я распознал в твоей корзине, ты можешь приготовить лазанью, греческий салат и глинтвейн.

Но тебе не хватает некоторых позиций. Добавить их?

}

Input на Classic Rec

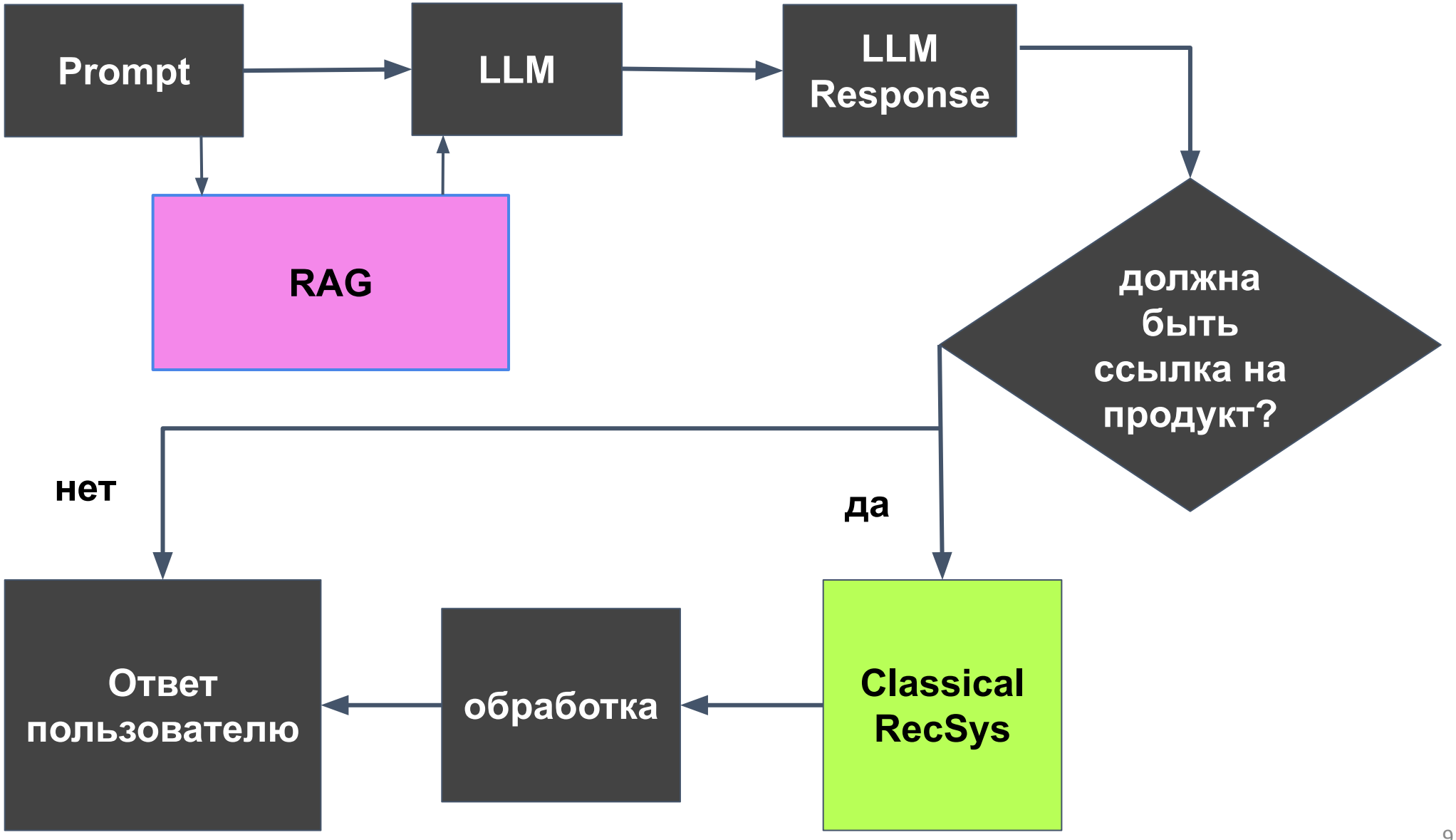
{

user_id: ‘id131313’,

products: {‘cheese_brie’, ‘honey’, ‘corn_bread’}

}

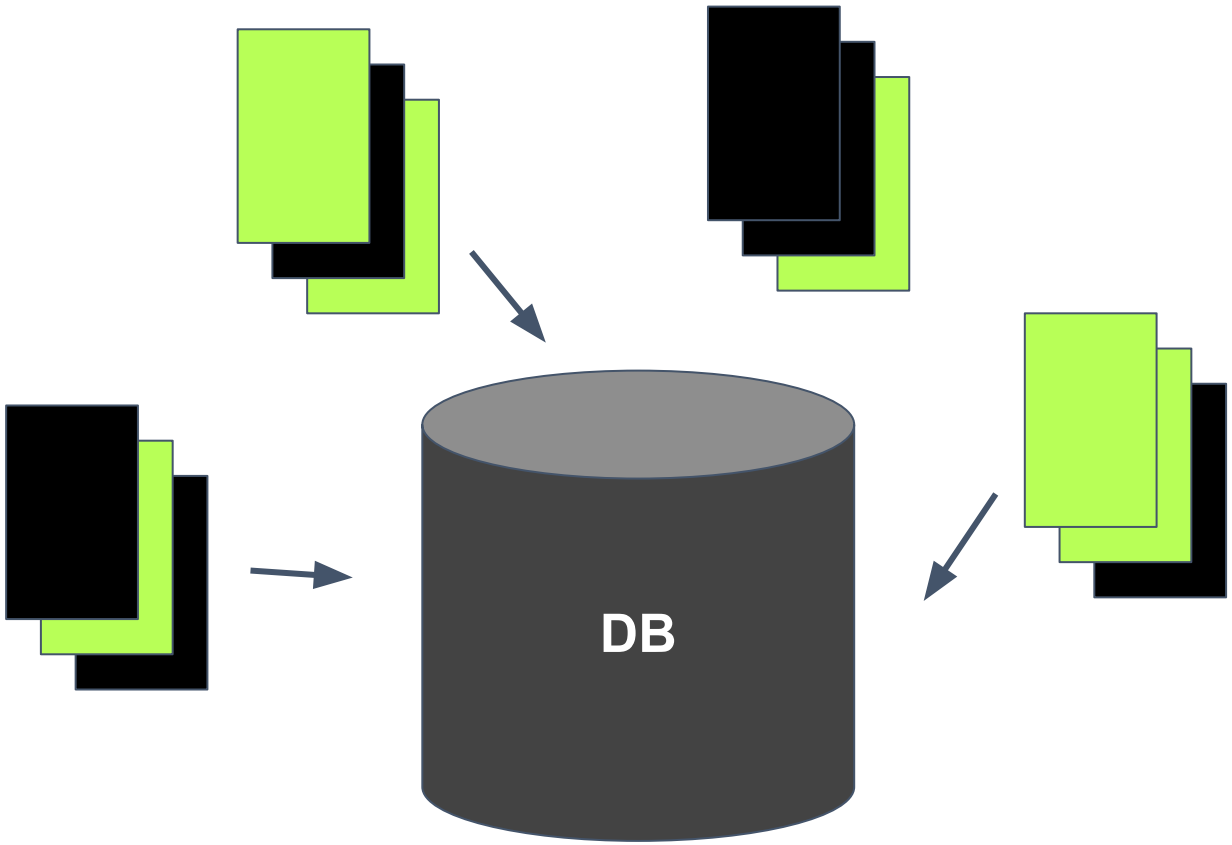
X5 Digital


Prompt
LLM
LLM Response
RAG
должна быть ссылка на продукт?
нет
да
Ответ пользователю
обработка
Classical RecSys

X5 Digital

ДАННЫЕ

X5 Digital



X5 Digital

Наборы данных

Тип данных	Порядок количества
Название, описания товаров, состав	5 000 - 40 000
Разбиение по стандартным категориям из каталожного дерева	до 50
Разбиения на data driven категории, вытащенные в процессе EDA	до 10 разбиений, в каждом 3-4 категории
Данные сценариев использования товаров - Датасеты рецептов	до 1000 различных уникальных* рецептов
Списки комплементарных товаров Формируем из датасетов рецептов (рецепты Впрок, датасеты Recipe Box и Recipe1M)	до 10 000
Персональные шаблоны пользователя Список нежелательных товаров / категорий товаров, список любимых товаров / категорий**	2

*имеется в виду уникальность после кластеризации и чистки неточных дублей на основе состава и способа приготовления
**эти пользовательские любимые категории могут сильно отличаться от категорий каталожного дерева

X5 Digital

Первичная подготовка данных

- Чистка текстов от мусора (регулярки)
  - вспомогательные и непечатные символы
  - разметка страницы
  - сокращения
  - ненужная техническая информация, сео-оптимизация и поисковые теги

X5 Digital

Первичная подготовка данных

- Чистка текстов от мусора (регулярки)
  - вспомогательные и непечатные символы
  - разметка страницы
  - сокращения
  - ненужная техническая информация, сео-оптимизация и поисковые теги

- Удаление дублирующейся информации в самих данных
  - разбиение на чанки
  - измерение контекстной близости чанков внутри одного документа

X5 Digital

Первичная подготовка данных

- Чистка текстов от мусора (регулярки)
  - вспомогательные и непечатные символы
  - разметка страницы
  - сокращения
  - ненужная техническая информация, сео-оптимизация и поисковые теги

Удаление дублирующейся информации в самих данных
  - разбиение на чанки
  - измерение контекстной близости чанков внутри одного документа

- Приведение рецептов к единым мерам измерения (первичный анализ мер и регулярки по замене)

X5 Digital

Первичная подготовка данных

- Чистка текстов от мусора (регулярки)
  - вспомогательные и непечатные символы
  - разметка страницы
  - сокращения
  - ненужная техническая информация, сео-оптимизация и поисковые теги

- Удаление дублирующейся информации в самих данных
  - разбиение на чанки
  - измерение контекстной близости чанков внутри одного документа
- Приведение рецептов к единым мерам измерения (первичный анализ мер и регулярки по замене)
- Если какие-то источники данных на другом языке, перевод, его валидация и чистка этих данных

X5 Digital

Убираем смысловые дубли

- Много данных – хорошо, но лучше, когда они уникальные в смысловом плане

X5 Digital

Убираем смысловые дубли

- Много данных – хорошо, но лучше, когда они уникальные в смысловом плане

- Убираем дубликаты рецептов (нечеткие дубликаты в смысле текстов)
  - смотрим относительно состава
  - могут быть незначительные модификации
  - смотрим относительно способа приготовления
  - могут быть одни и те же продукты, приготовленные разными путями – это разные рецепты

X5 Digital

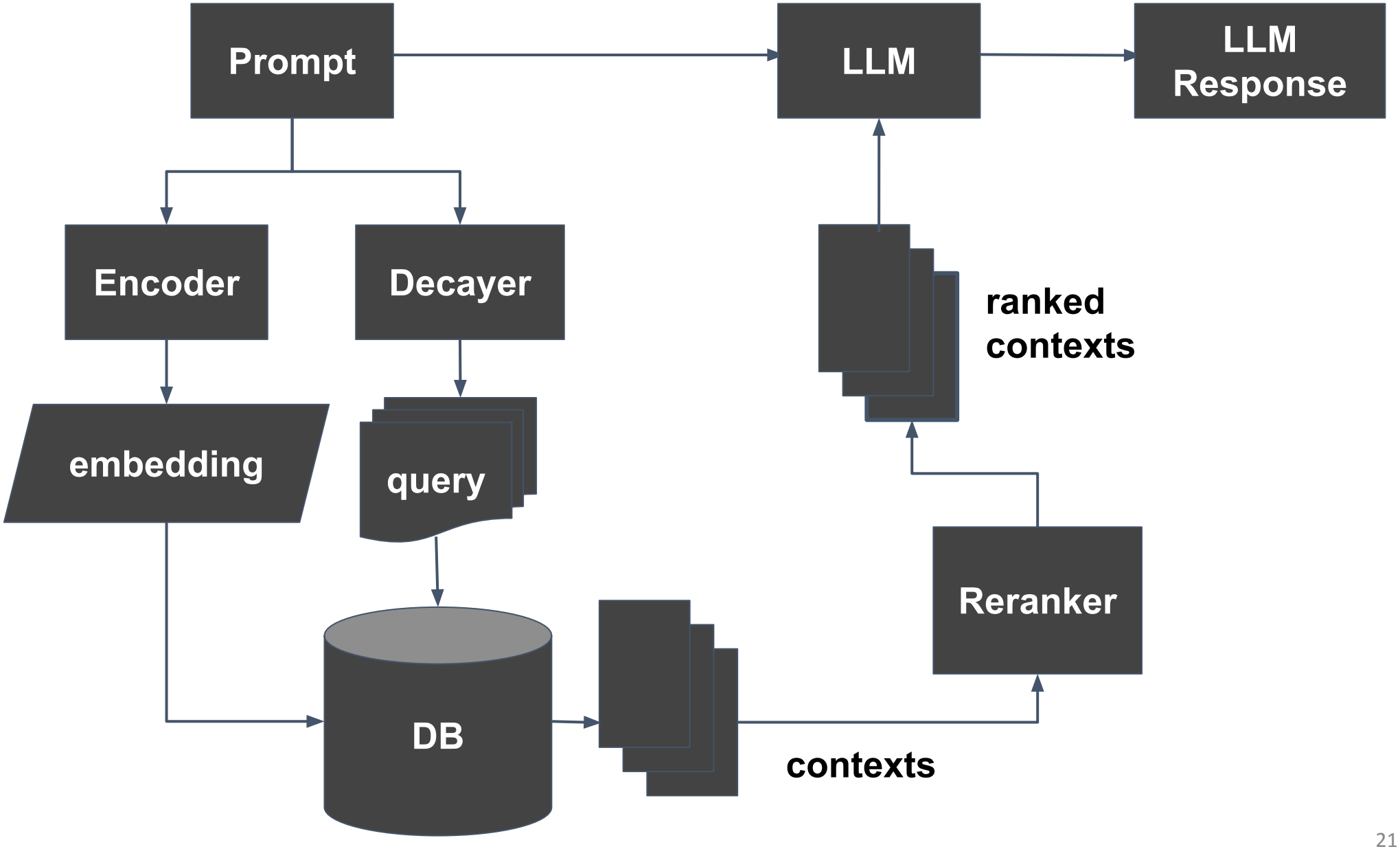
Убираем смысловые дубли

- Много данных – хорошо, но лучше, когда они уникальные в смысловом плане
- Убираем дубликаты рецептов (нечеткие дубликаты в смысле текстов)
  - смотрим относительно состава
  - могут быть незначительные модификации
  - смотрим относительно способа приготовления
  - могут быть одни и те же продукты, приготовленные разными путями – это разные рецепты
- Разбиение рецептов на чанки текста, измерение контекстной близости отрезков между датасетом
  - косинусной мерой
  - бертом

X5 Digital

Теперь организуем крутой RAG

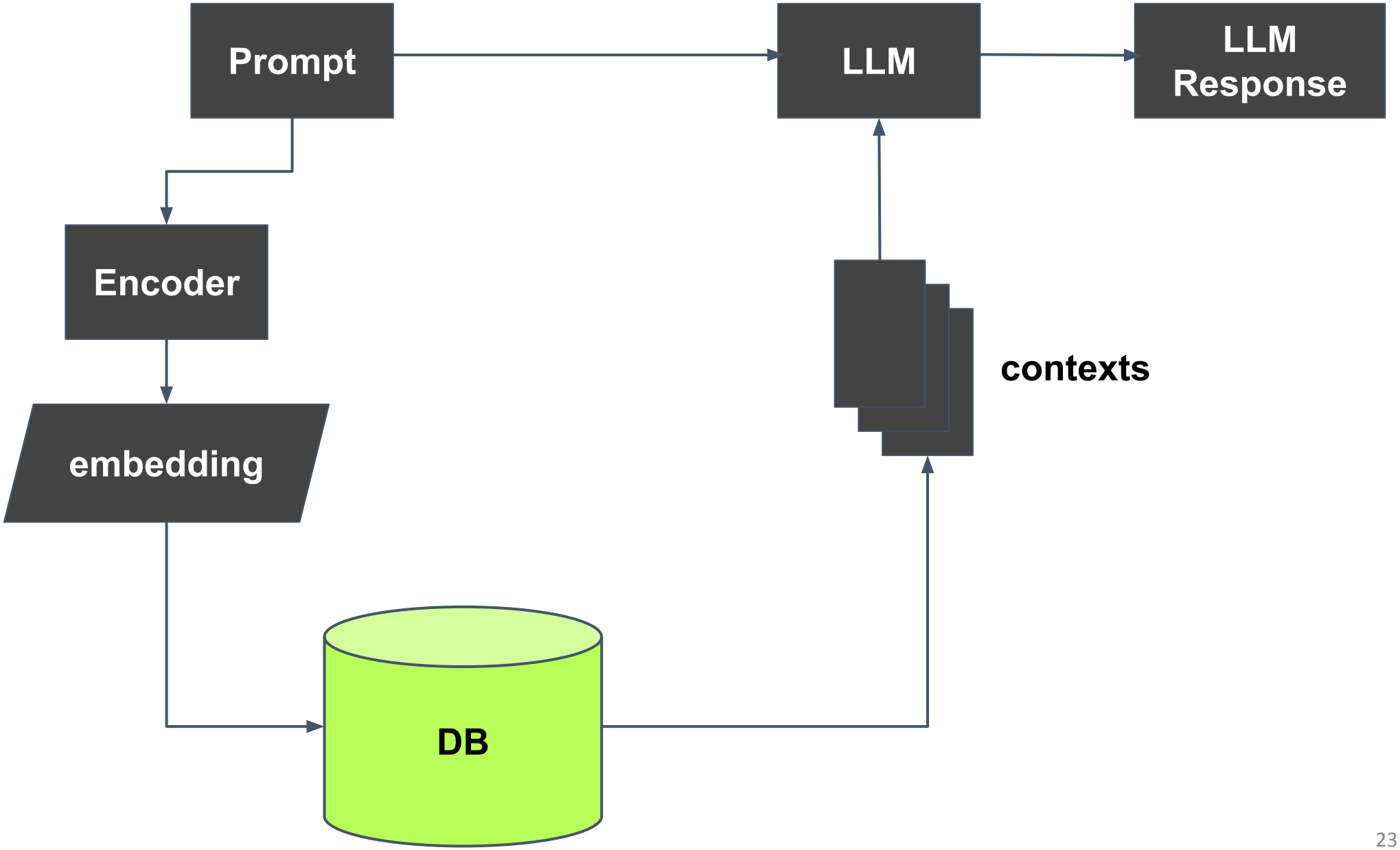
X5 Digital


Prompt
LLM
LLM Response
Encoder
Decayer
ranked contexts
embedding
query
Reranker
DB
contexts

X5 Digital

Векторные базы данных для RAG

X5 Digital

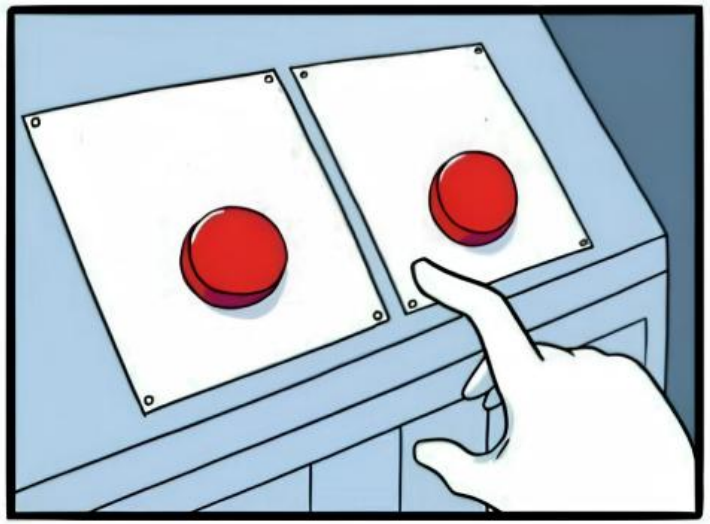


X5 Digital

Векторные базы данных

- Elastic Search
- Pinecone
- Chroma
- Qdrant
- PG Vectors над PostgreSQL

и т.д.



какую выбрать?

X5 Digital

Векторные базы данных

	Pinecone	Weaviate	Milvus	Qdrant	Chroma	Elasticsearch	PGvector
Is open source	❌	✅	✅	✅	✅	❌	✅
Self-host	❌	✅	✅	✅	✅	✅	✅
Cloud management	✅	✅	✅	✅	❌	✅	(✔)
Purpose-built for Vectors	✅	✅	✅	✅	✅	❌	❌
Developer experience	👍👍👍	👍👍	👍👍	👍👍	👍👍	👍	👍
Community	Community page & events	8k☆ github, 4k slack	23k☆ github, 4k slack	13k☆ github, 3k discord	9k☆ github, 6k discord	23k slack	6k☆ github
Queries per second (using text nytimes-256-angular)	150 *for p2, but more pods can be added	791	2406	326	?	700-100 *from various reports	141
Latency, ms (Recall/Percentile 95 (millis), nytimes-256-angular)	1 *batched search, 0.99 recall, 200k SBERT	2	1	4	?	?	8
Supported index types	?	HNSW	Multiple (11 total)	HNSW	HNSW	HNSW	HNSW/IVFFlat

https://zackproser.com/blog/vector-databases-compared

https://benchmark.vectorview.ai/vectordbs.html

X5 Digital

Всегда ли нужно срочно принимать это решение?

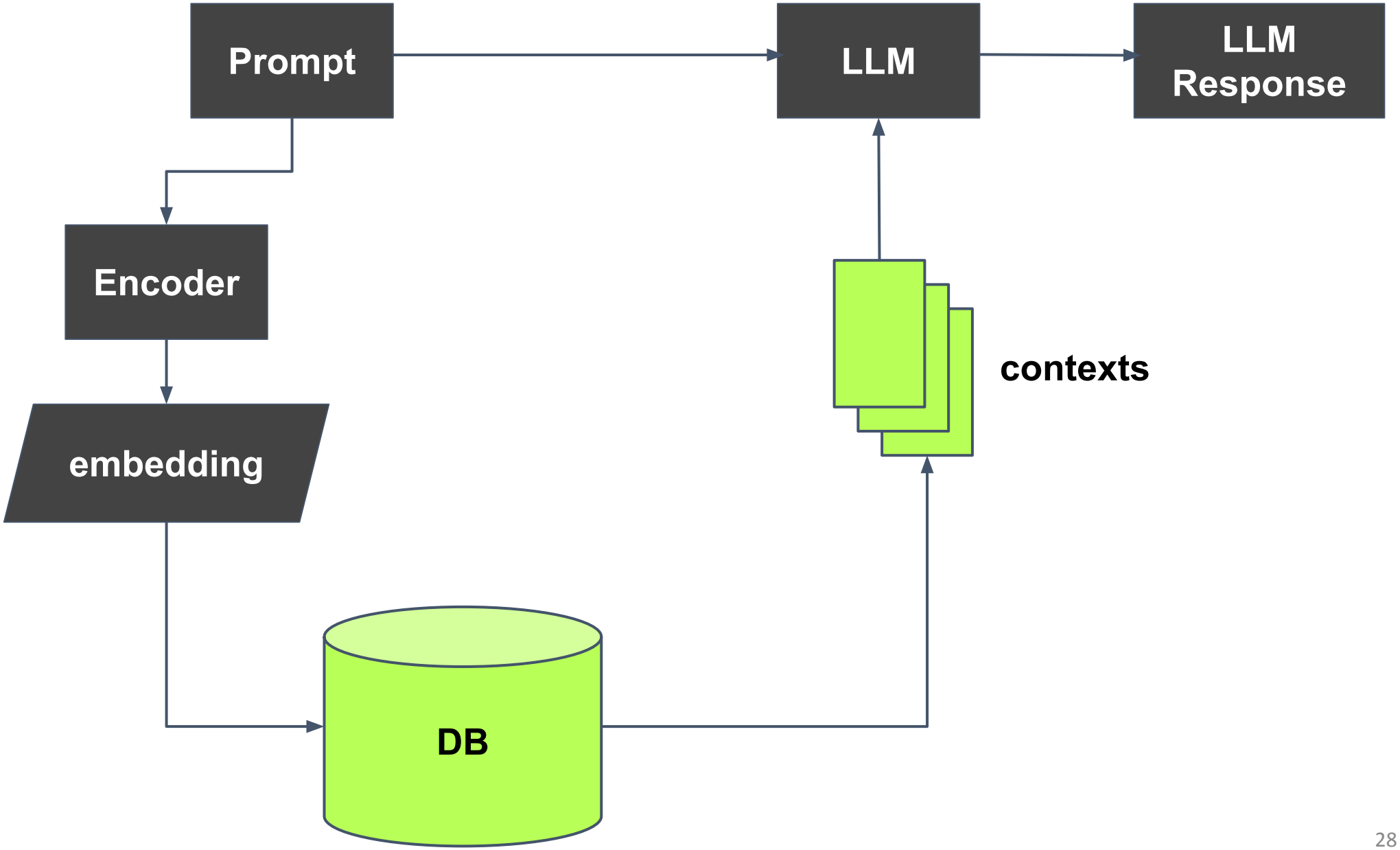
- **Библиотеки векторного поиска**
  Faiss
  Annoy
  HNSWLib

  Разница с векторной базой
  либа работает со статическими данными
  в БД реализованы операции CRUD
  либа для прототипирования хороша

X5 Digital

Окей, делаем векторный поиск

X5 Digital



X5 Digital

Как искать релевантные документы?

1

Выберем и построим индекс

2

Найдём кучу близких по смыслу документов

3

Отранжируем их

X5 Digital

Как искать релевантные документы?

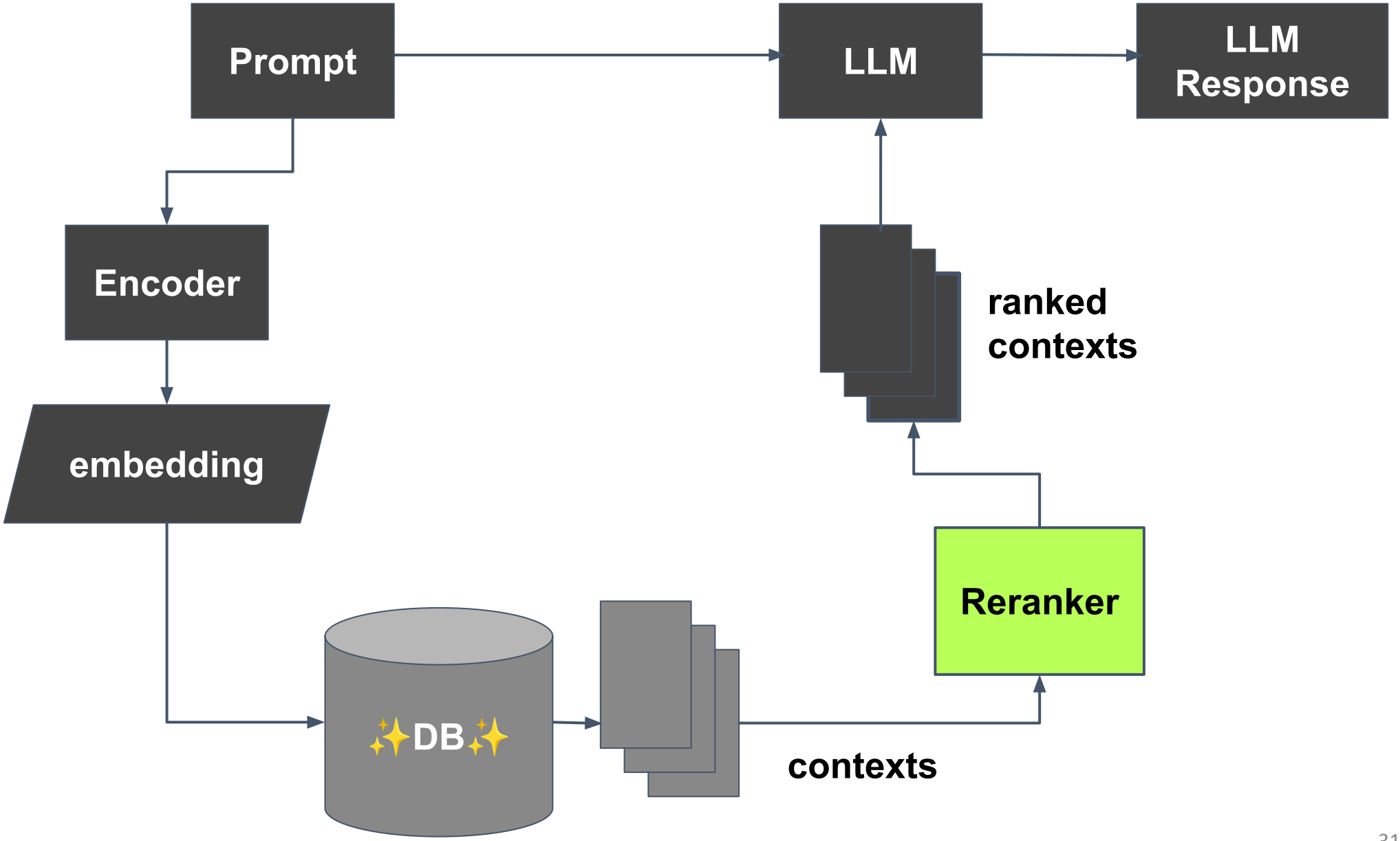
1. построение индексов
   - HNSW
   - IndexIVFPQ, etc.

   или другие

2. поиск кучи контекстно близких документов
   - метрическая задача на структуре индекса

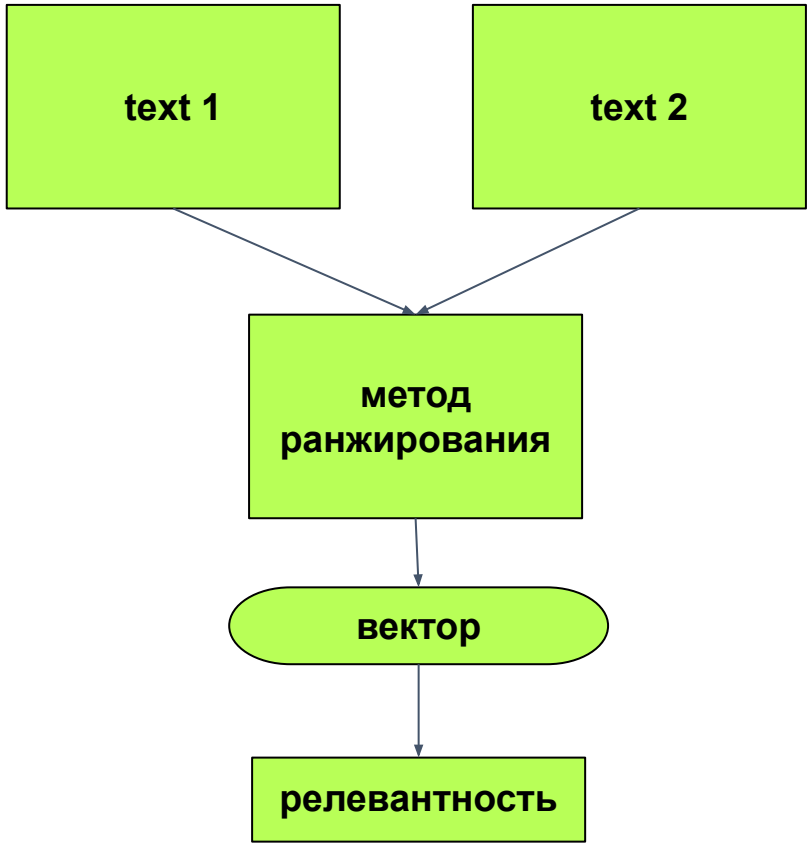
3. а как будем ранжировать?

X5 Digital


Prompt
LLM
LLM Response
Encoder
ranked contexts
embedding
Reranker
✨DB✨
contexts

X5 Digital

Методы определения релевантности



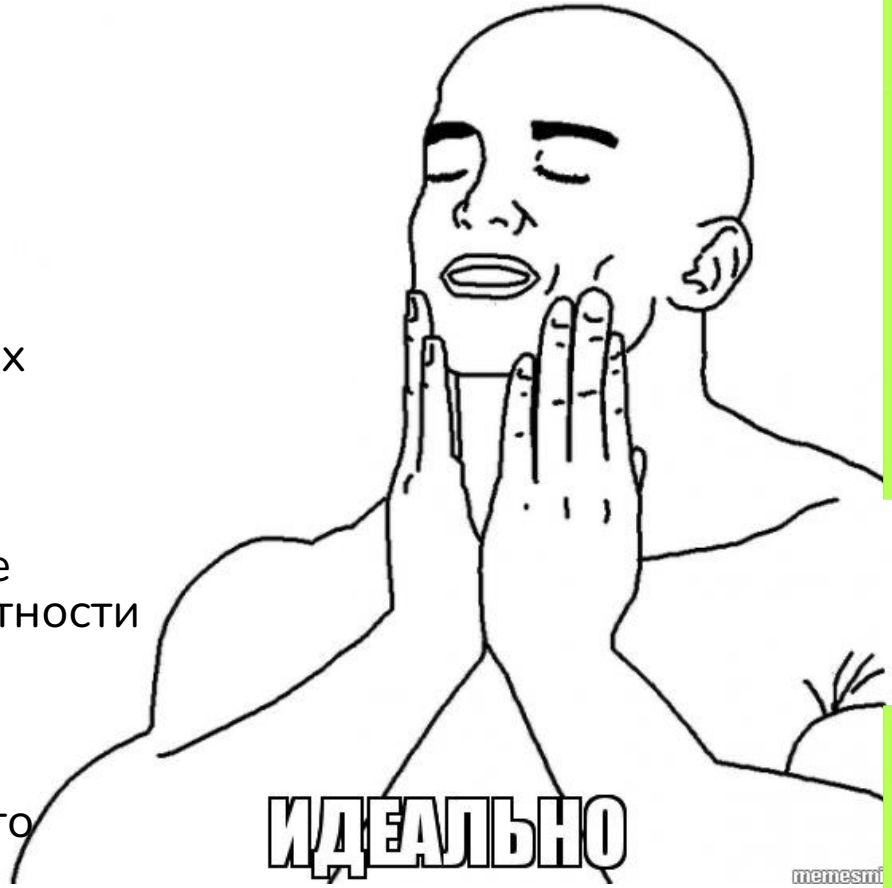
- Bert Cross Encoder или другие трансформеры
- Reciprocal Rank Fusion
- TF-IDF, BestMatch-25

можно придумывать свои кастомные способы оценить релевантность документов, добавляя коэффициенты, модифицируя формулы, собирая ансамбли

X5 Digital

Как оценить качество ранжирования

- P@K
  - без учета порядка, для отдельно взятого объекта
- MAP@K
  - можно применять для кучи пользователей / поисковых запросов, считаем ap@k для каждого и усредняем
- NDCG@K
  - чем больше релевантных элементов в топе, тем лучше
  - учитывает порядок элементов домножением релевантности на вес
- MRR
  - средний обратный ранг первого правильно угаданного элемента



https://education.yandex.ru/handbook/ml/article/zadacha-ranzhirovaniya

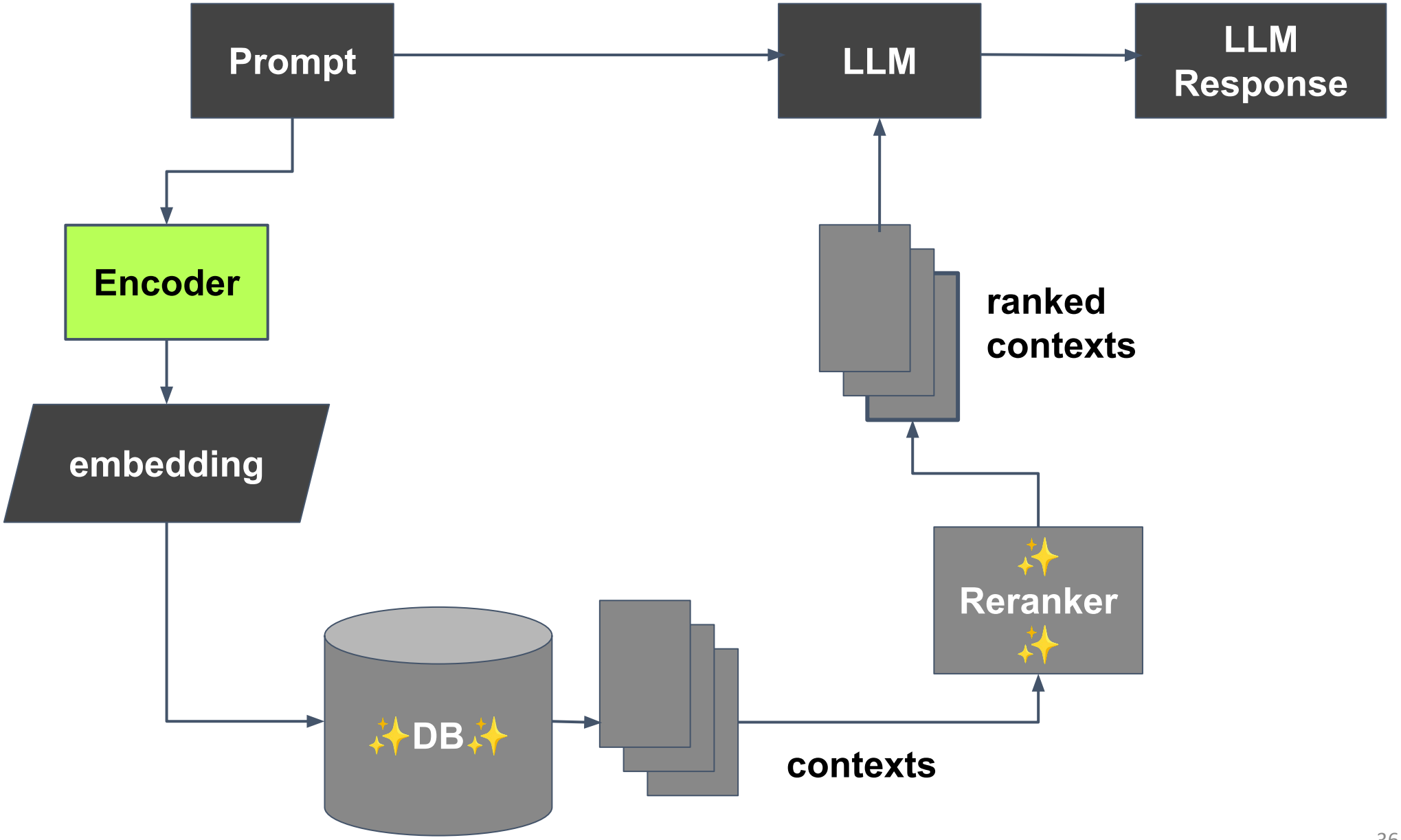
X5 Digital

Хотим еще кастомнее

X5 Digital

(кастомизируем вектора в базе)

X5 Digital

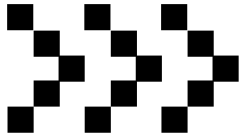

Prompt
LLM
LLM Response
Encoder
ranked contexts
embedding
Reranker
DB
contexts

X5 Digital

Metric Learning

Методы:

- Additive Margin Softmax
- PLDA
- Различные методы нормализации векторов, разносящие / сближающие их в пространстве
- Triplet / contrastive loss-ы



Хотим пространство, где похожие друг на друга находятся очень близко,

а различающиеся - как можно больше далеко

X5 Digital

Особенности разбиения текстов перед векторизацией

- Чем меньше текст по размеру, тем меньше там различных смыслов, тем лучше будет работать контекстный поиск
- Разбиение текстов на подтексты должны немного перекрывать друг друга, они должны быть последовательностью связных смыслов
- Внутри подтекстов не должно быть разрывов мысли

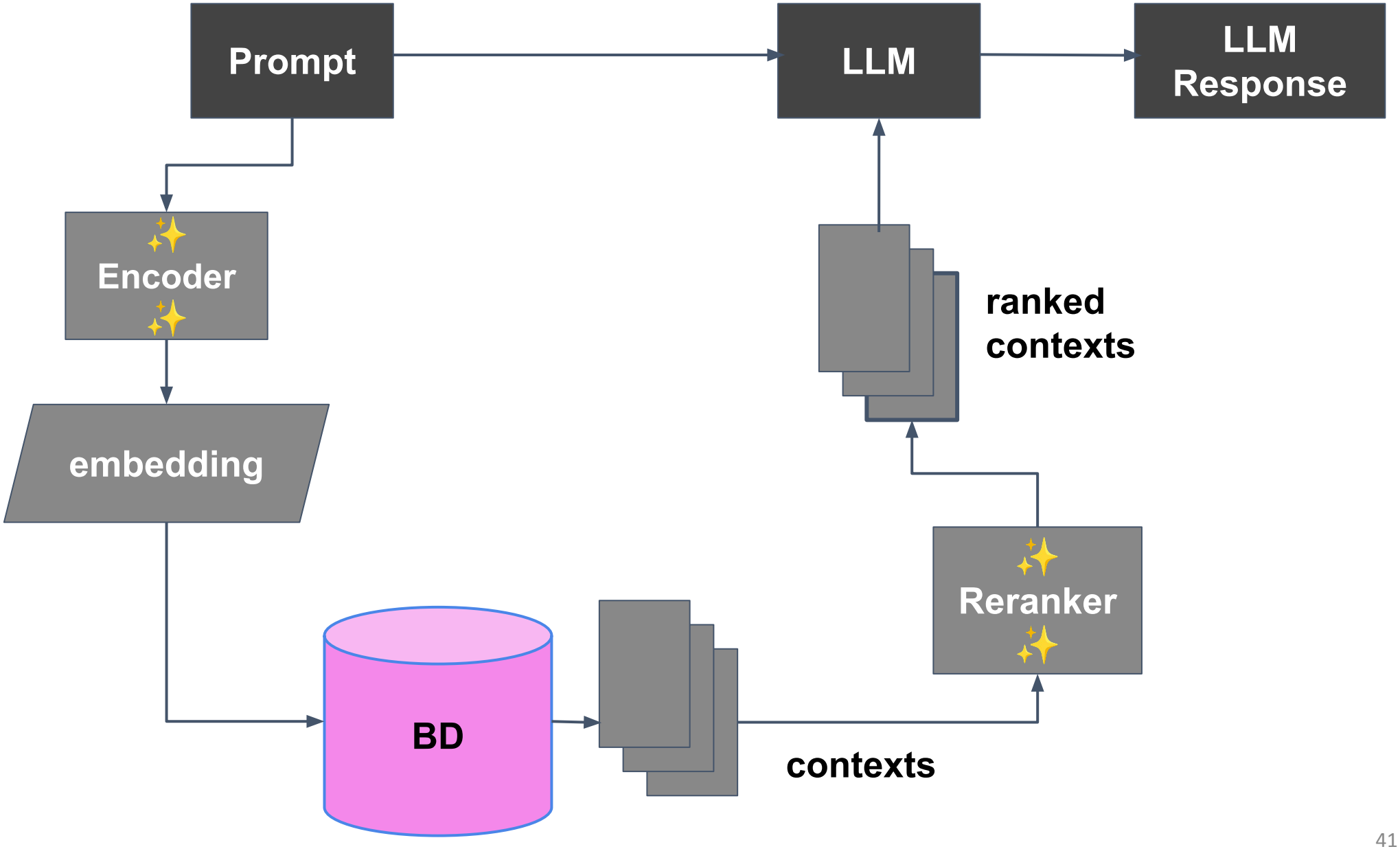
X5 Digital

Можно ли еще что-то поменять?

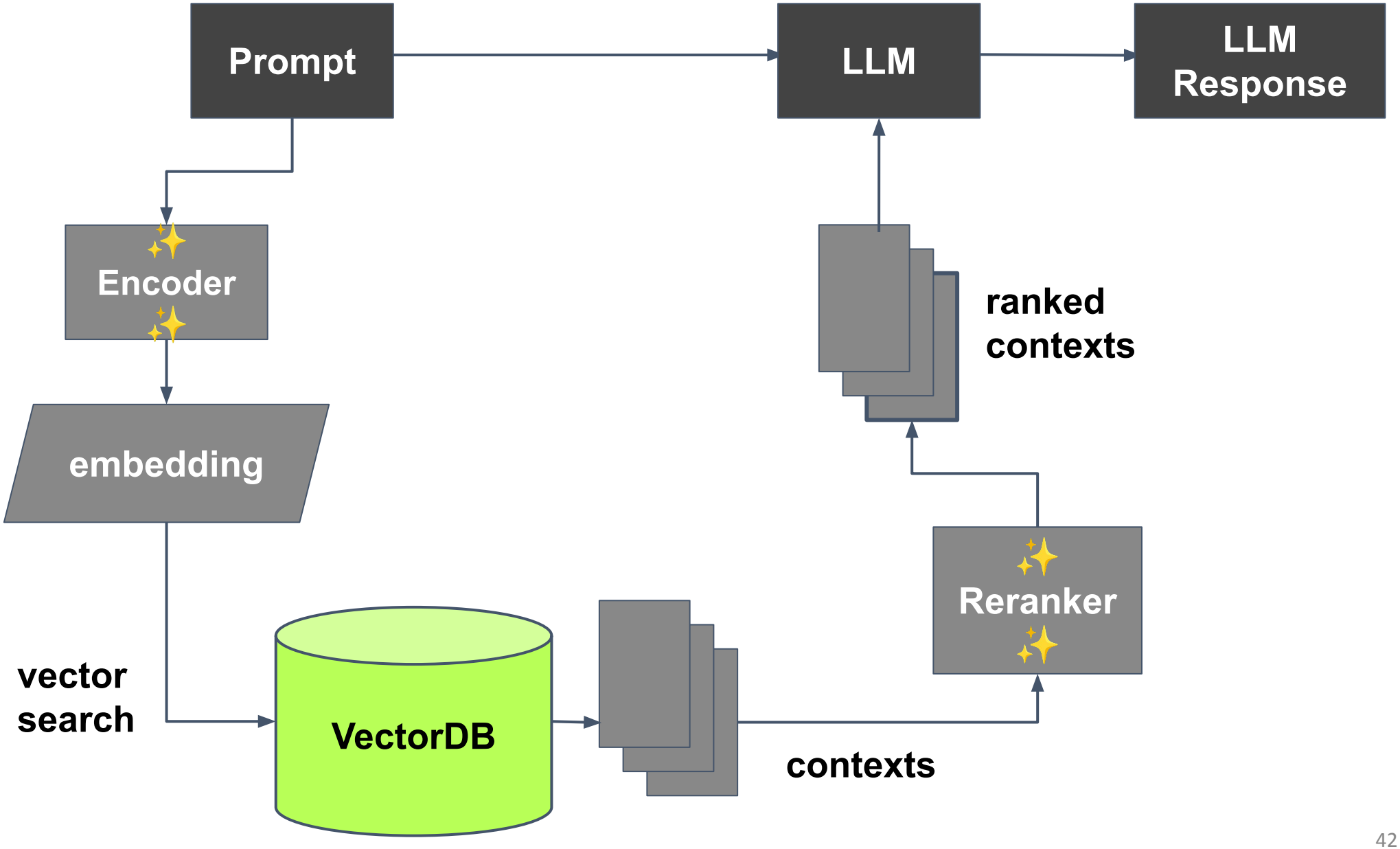
X5 Digital

Да 😏

X5 Digital


Prompt
LLM
LLM Response
Encoder
embedding
ranked contexts
Reranker
BD
contexts

X5 Digital


Prompt
LLM
LLM Response
Encoder
embedding
ranked contexts
Reranker
vector search
VectorDB
contexts

X5 Digital

RAG с векторной базой данных

- влияем на полноту поиска только косвенно

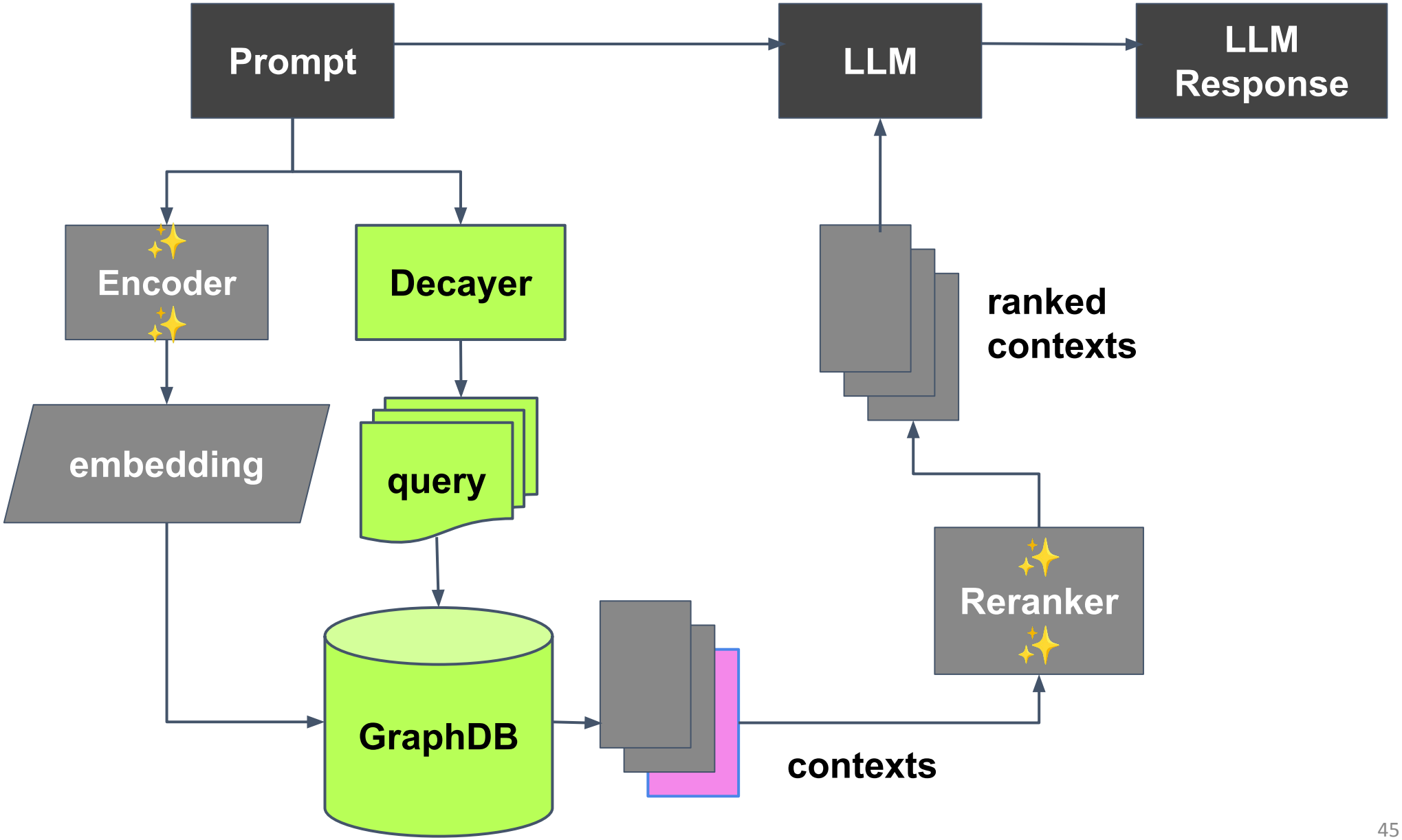
X5 Digital

RAG с Neo4J

✓ оставляем векторный поиск

✓ делаем еще и графовый поиск параллельно

X5 Digital


Prompt
LLM
LLM Response
Encoder
Decayer
ranked contexts
embedding
query
Reranker
GraphDB
contexts

X5 Digital

Выстраиваем из данных граф

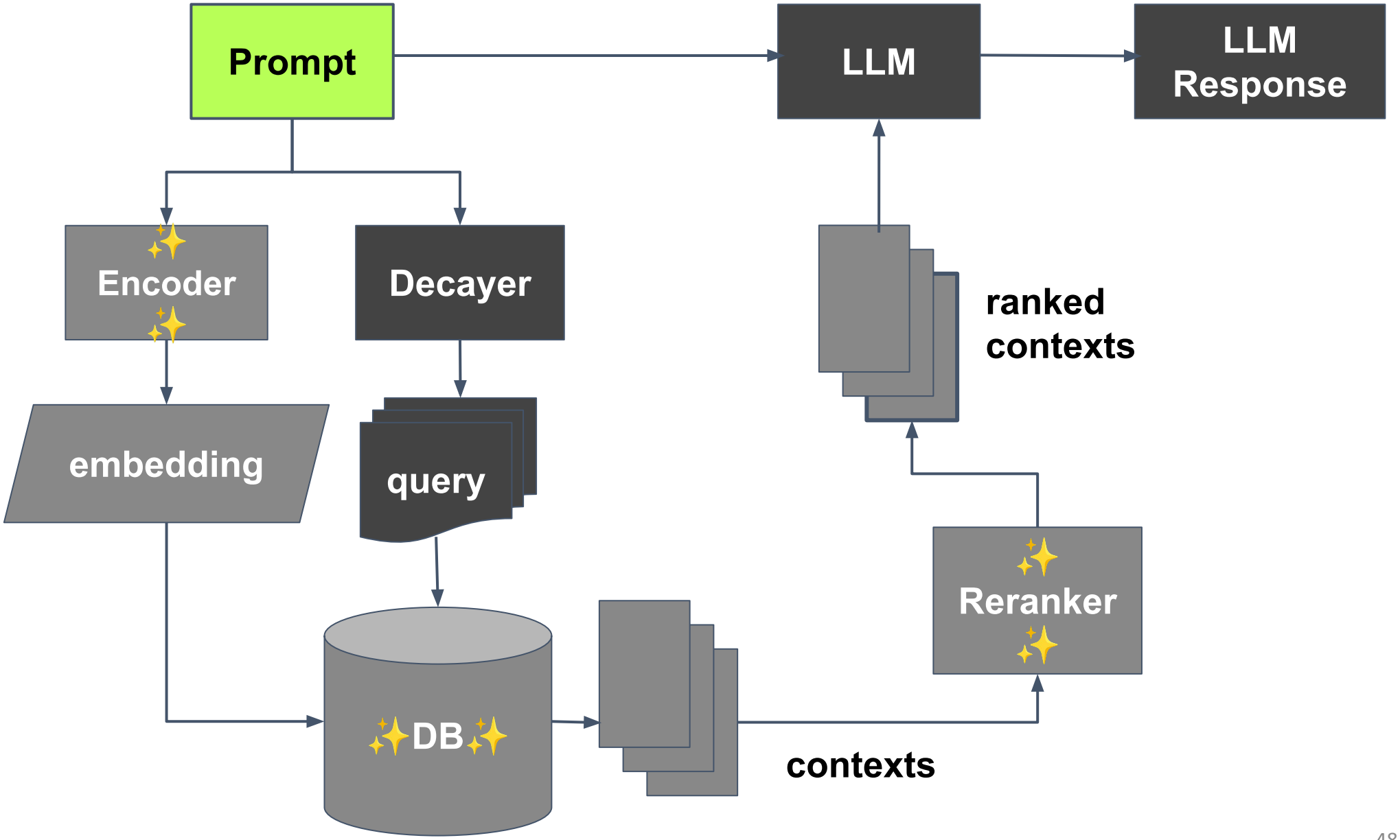
связываем

- продукты с рецептами,
- продукты с составами / КБЖУ,
- продукты со стандартными и кастомными категориями

X5 Digital

Еще можно кастомизировать сам промпт

X5 Digital


Prompt
LLM
LLM Response
Encoder
Decayer
ranked contexts
embedding
query
Reranker
DB
contexts

X5 Digital

Аугментируем входящую информацию

- Первоначальный запрос пользователя перефразируем до 5 раз через LLM
- По каждому из них ищем чанки, затем суммаризируем их
- Если найдено множество совпадений, также суммаризируем вначале

X5 Digital

Добавляем инструкции от себя

можно добавить инструкции в промпт после получения документов из базы

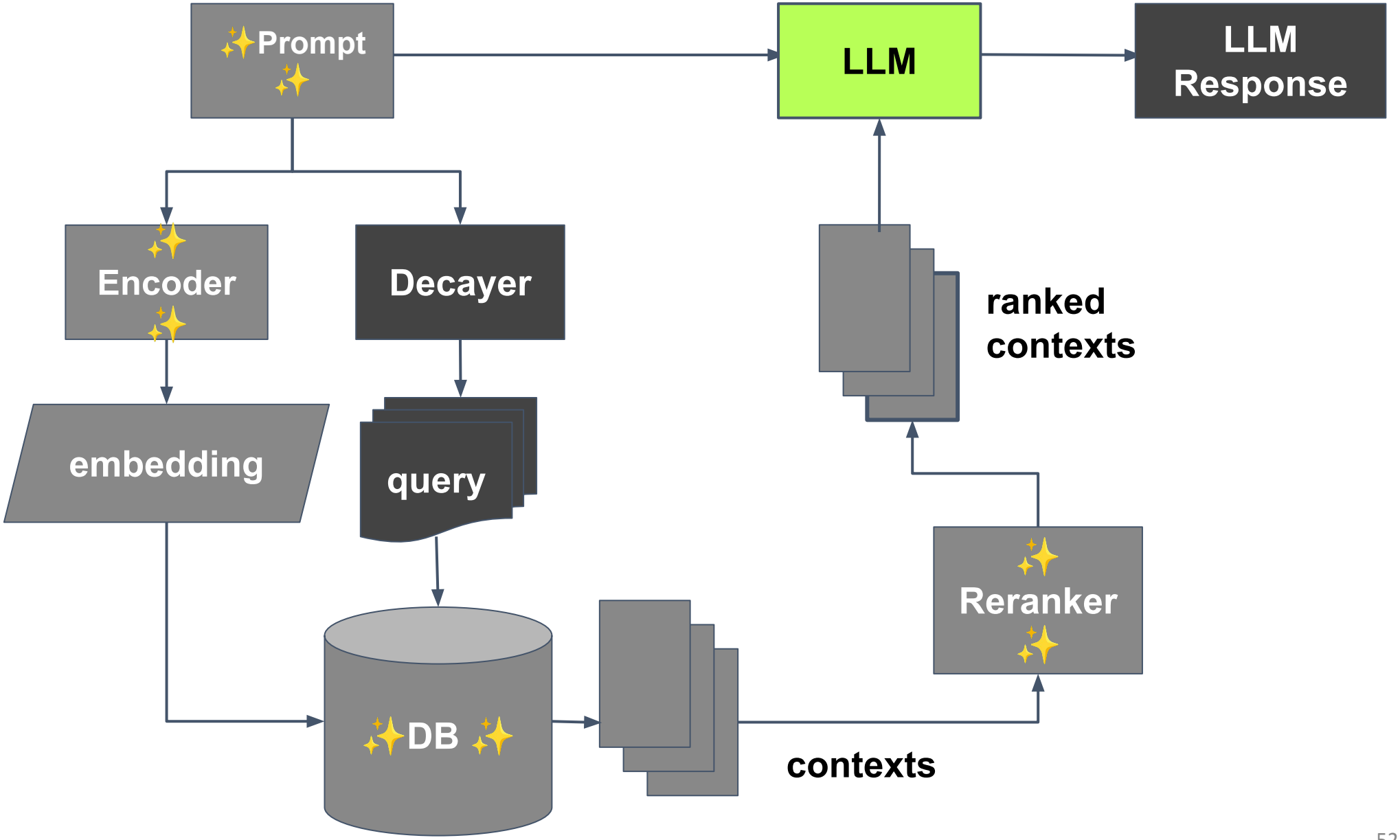
- хард-кодим инструкции для нужных нам юзкейсов

- добавляем эмпирически наблюдаемые костыли типа
  - не спеши, внимательно подумай
  - обещаем модели деньги
  - и т.д.

X5 Digital

Только RAG?
А модель учить будем?

X5 Digital


Prompt
LLM
LLM Response
Encoder
Decayer
ranked contexts
embedding
query
Reranker
DB
contexts

X5 Digital

PEFT на своем домене данных

- улучшает результат на не сильно отличающимся от общего домене знаний
- от 100 примеров (против гигабайтов текстов на классическом файн тюне)
- не требуется каких-то значительных изменений модели мира внутри языковой модели
- использование легковесных адаптеров (в сотни-тысячи раз меньше основной модели)
- несколько gpu-часов (против сотен - тысяч на классическом файн тюне)

- **LoRA**
- **DoRA**
  - улучшенная модификация лоры
  - сильно превосходит оригинал
- **GaLore**
  - очень сильно оптимизированная по потребляемой памяти по сравнению с лорой
  - можно обучать большие модели на меньших видеокартах
- **P-tune / P-tune v2 / Prefix tuning**
  - устаревшие, LoRA обычно сильно лучше

X5 Digital

Дообучение модели до формата RAG

- Чтобы модель лучше понимала, что от нее требуется, можно ее дообучить на выбранный нами формат взаимодействия

X5 Digital

Квантизация инференса

X5 Digital

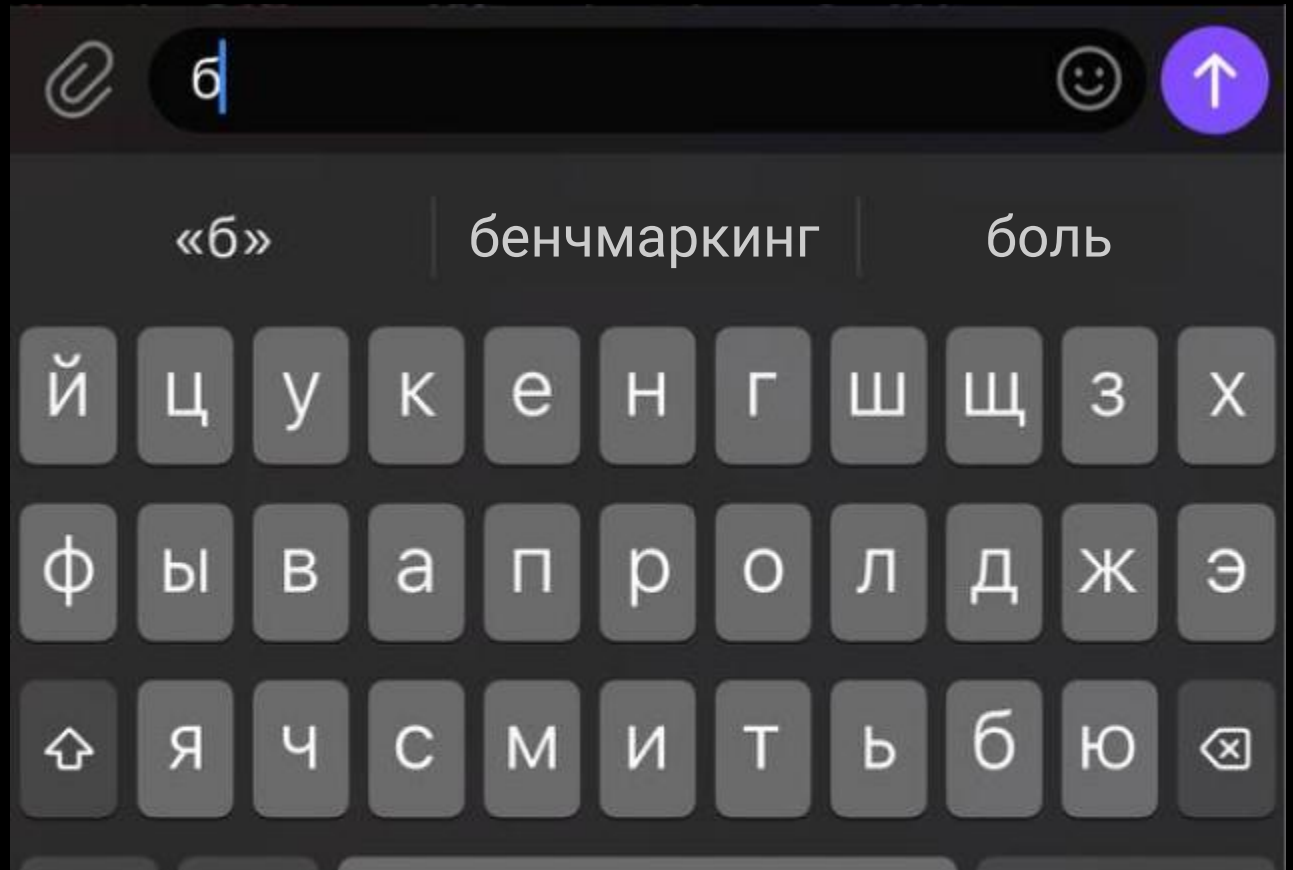
Квантизация FP16 в FP8 уменьшает говорливость

- уменьшается вероятность галлюцинаций
- появляется краткость и ясность изложения
- уходят часто ненужные детали
- уменьшаются хождения смысла вокруг да около
- более легко воспринимаются тексты в приложении

*на основе нашего опыта
** в FP4 уже часто получается бред, так что тоже нужно найти свой баланс

X5 Digital

Бенчмаркинг



X5 Digital

Классические открытые бенчи для LLM

- GLUE
- MMLU
- SQuAD
- MT-Bench
- HumanEval
- HELLASWAG
- WinoGrande
- BBQ
- TruthfulQA
- MLPerf

X5 Digital

Какая проблема с открытыми?

- все внешние бенчмарки до 80% утекают в обучающую выборку открытых моделей, поэтому они не показывают объективную картину
- есть Бот Арена – более объективно на общих задачах
- а специфические?

X5 Digital

Открытые специализированные бенчи по питанию

- TruthfulQA Nutrition
- HellaSwag.Food and Entertaining
- MMLU.Nutrition
- NutriBench

Формат заданий в них:

- выбрать правильный ответ из вариантов
- сравнить текстовые ответы модели и эталона
- оценить рассчитанное кол-во углеводов / белков / жиров / калорий

X5 Digital

Внутренние бенчмарки - это важно!

- Внутренние бенчи аналогично открытым, но со своими товарами

- Бенчи на понимание количественных мер

- Бенчи на знание базовой математики КБЖУ и ограничений по стоимости товаров

Размер: от 1000 кейсов

X5 Digital

Давайте оставаться на связи!

Мой канал
про data science

Telegram

LinkedIn