

RAG

DataFest 2024

Ai
Run

какая задача стоит?

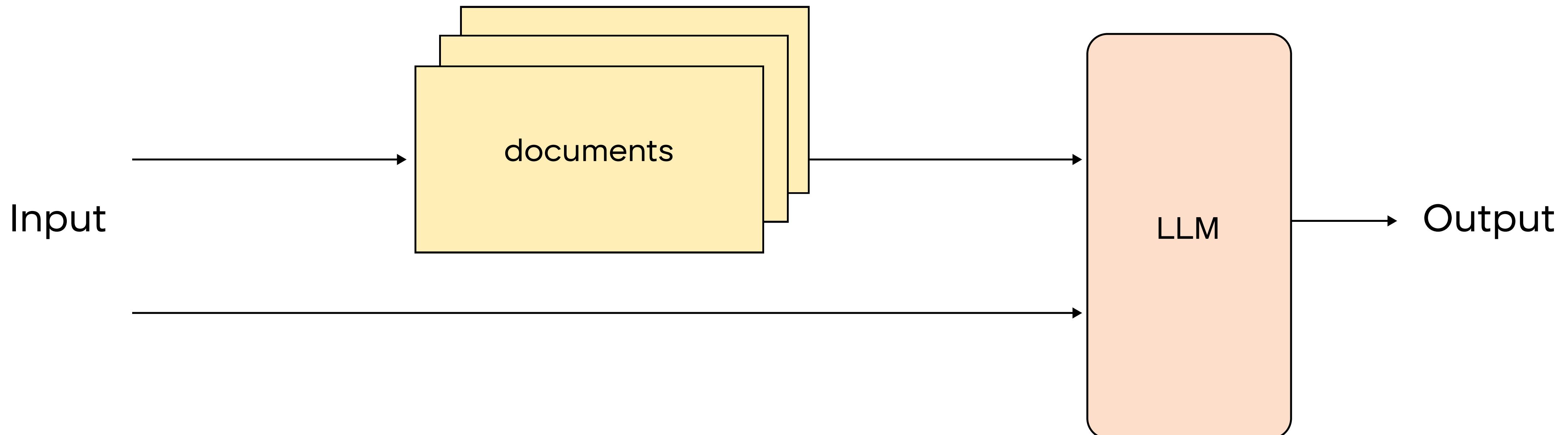
хотим внедрить rag в
корпоративные системы – что
использовать?

1

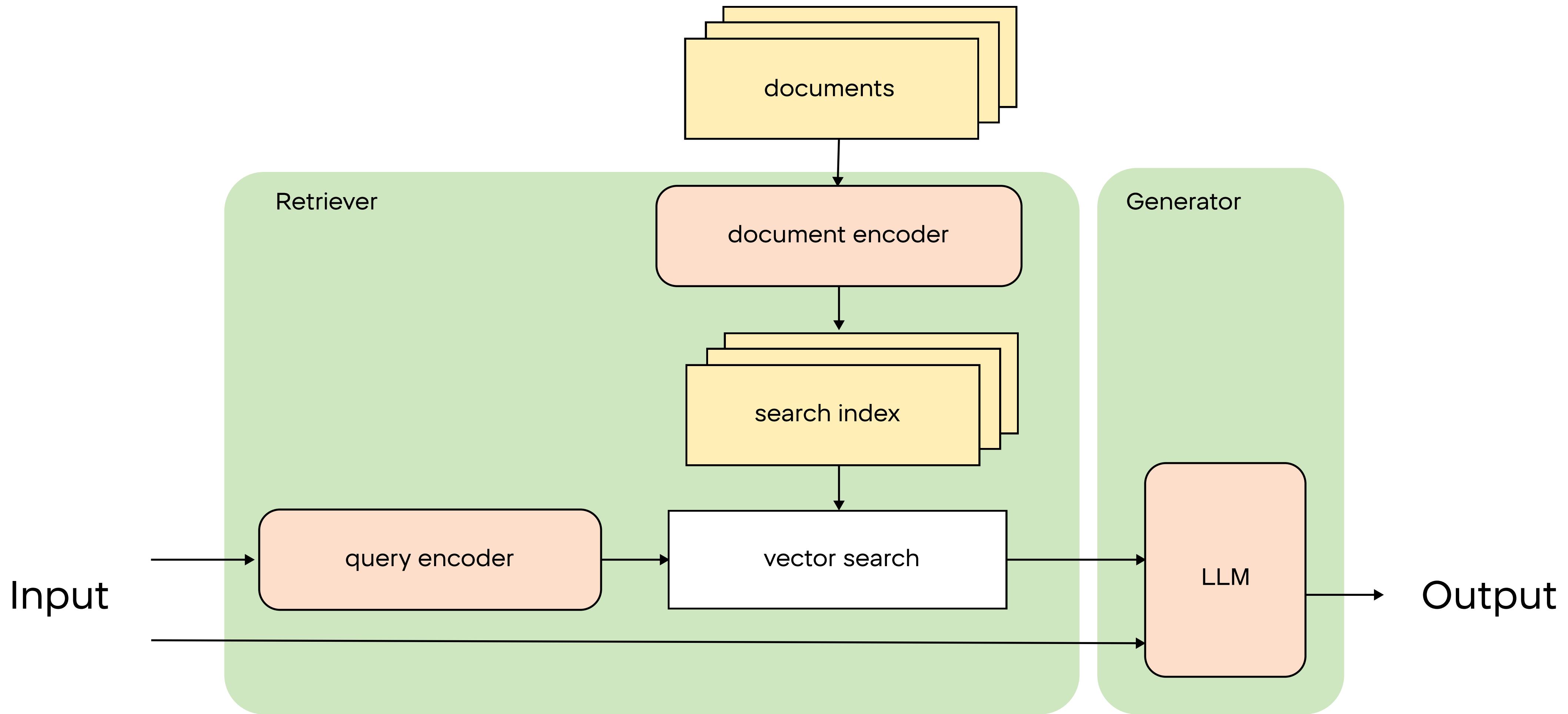
ЧТО ЗА МАГИЯ?

RAG – RETRIEVAL AUGMENTED GENERATION

RAG: просто

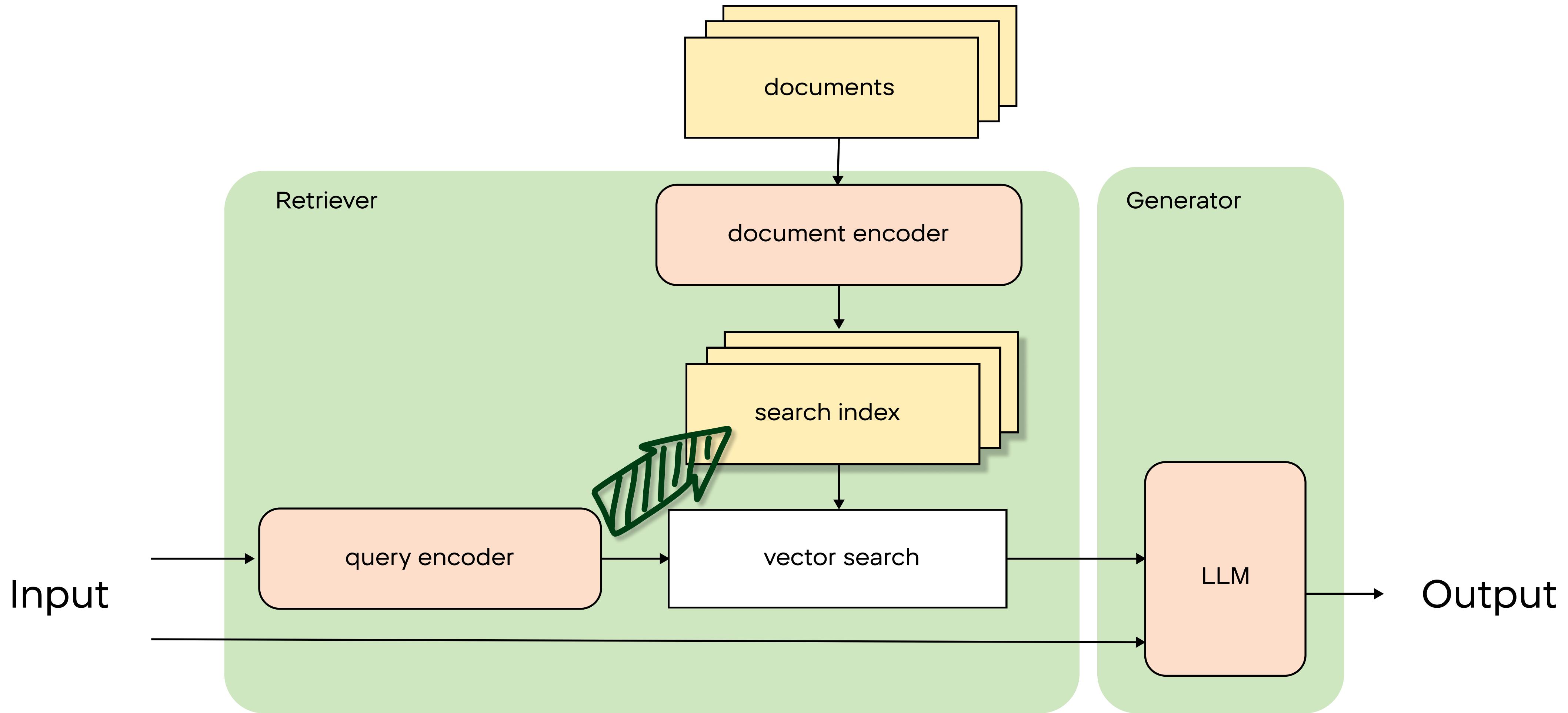


RAG: посложнее

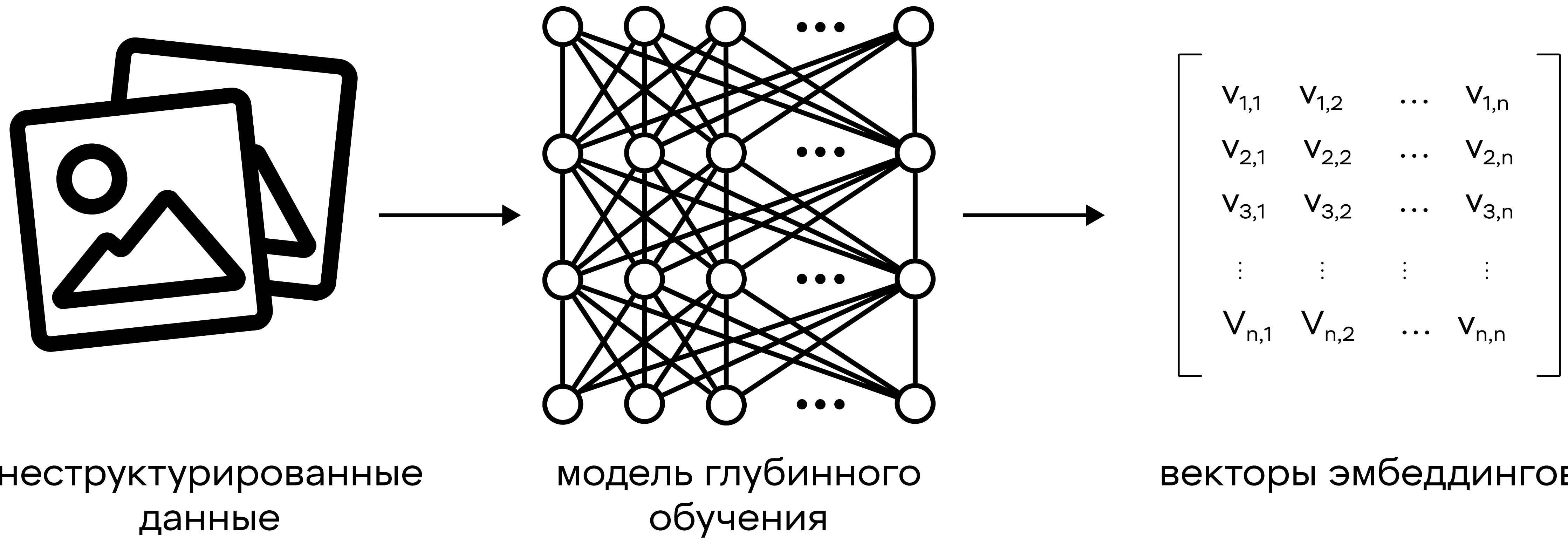


ВЕКТОРНЫЕ БАЗЫ ДАННЫХ

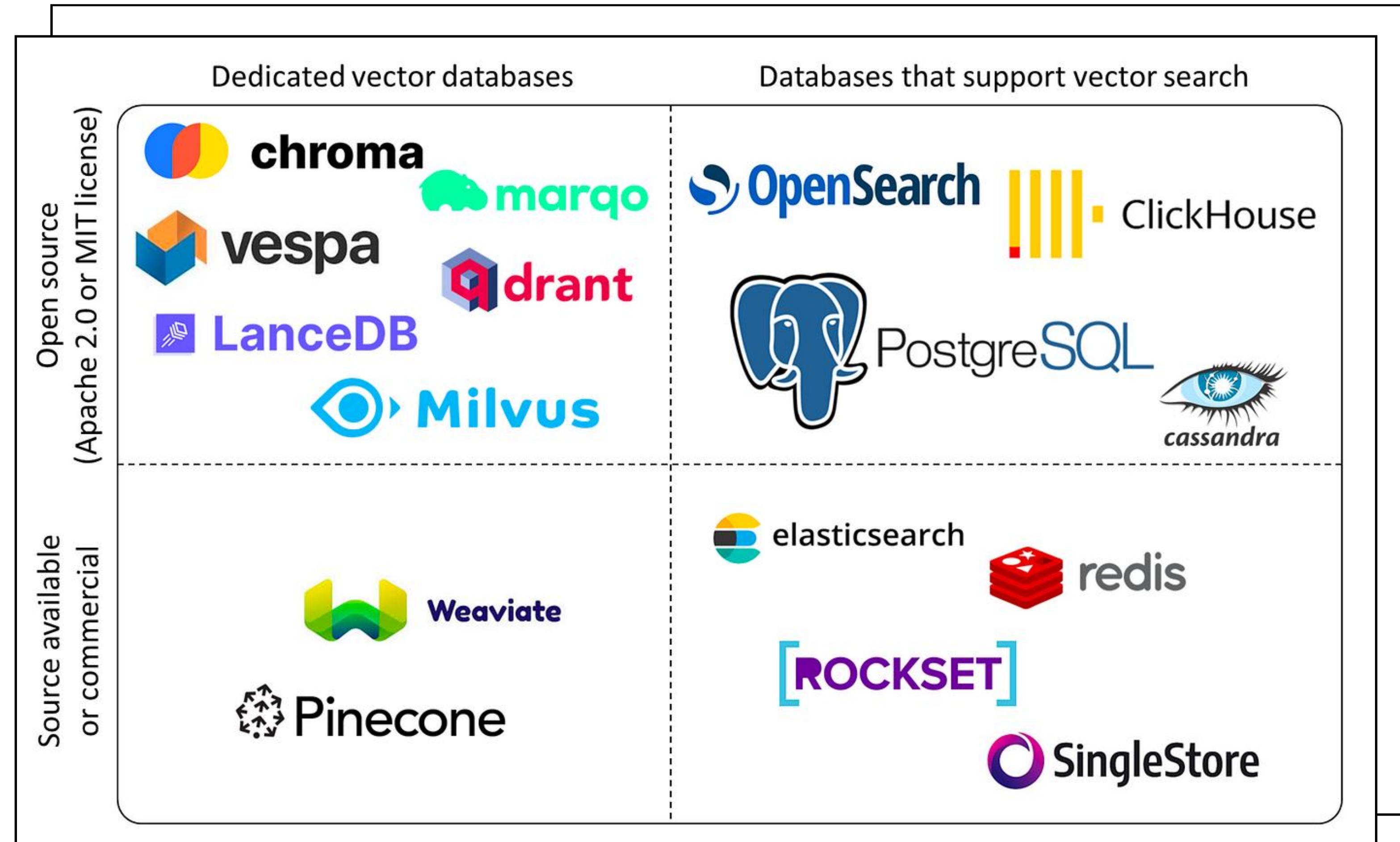
векторные базы данных



векторные базы данных: что это?



векторные базы данных: что это?

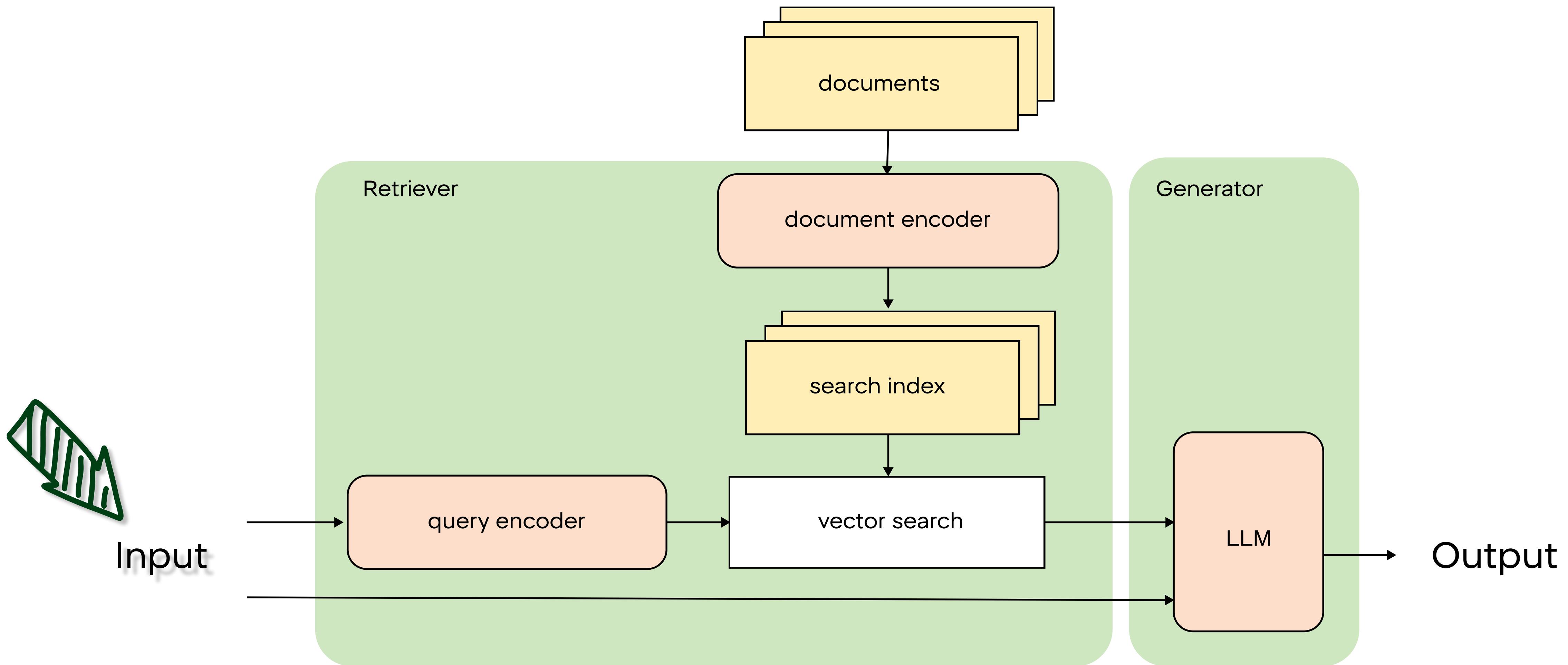


2

КАК ИСКАТЬ?

QUERY EXPANSION

query expansion

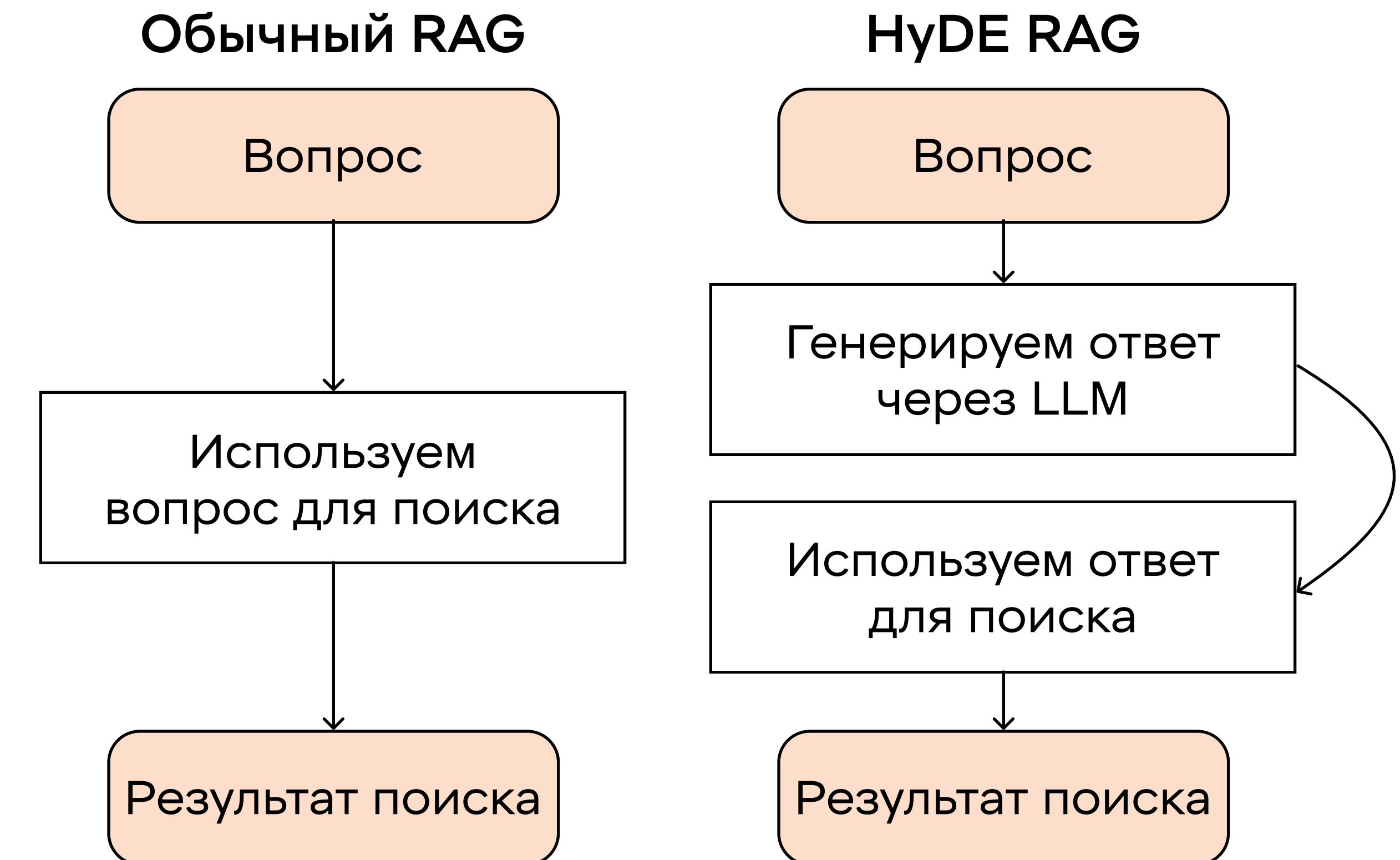


HyDE

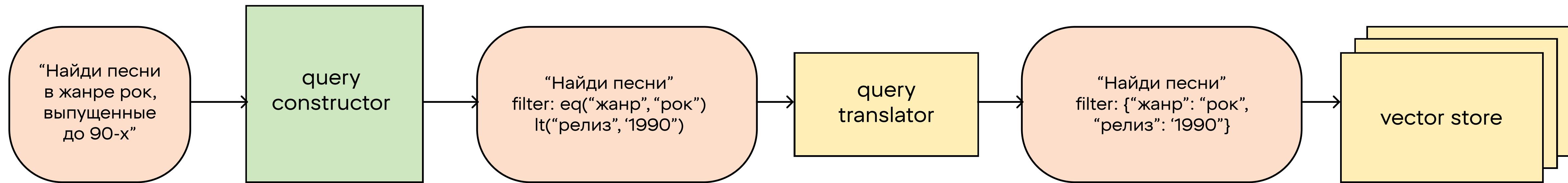
HYPOTHETICAL DOCUMENT EMBEDDINGS

идея:

- хотим потенциальным ответом уточнить наш запрос и расширить область поиска
- полезно, если хотим “раздвинуть” два очень похожих вопроса



SelfQuery

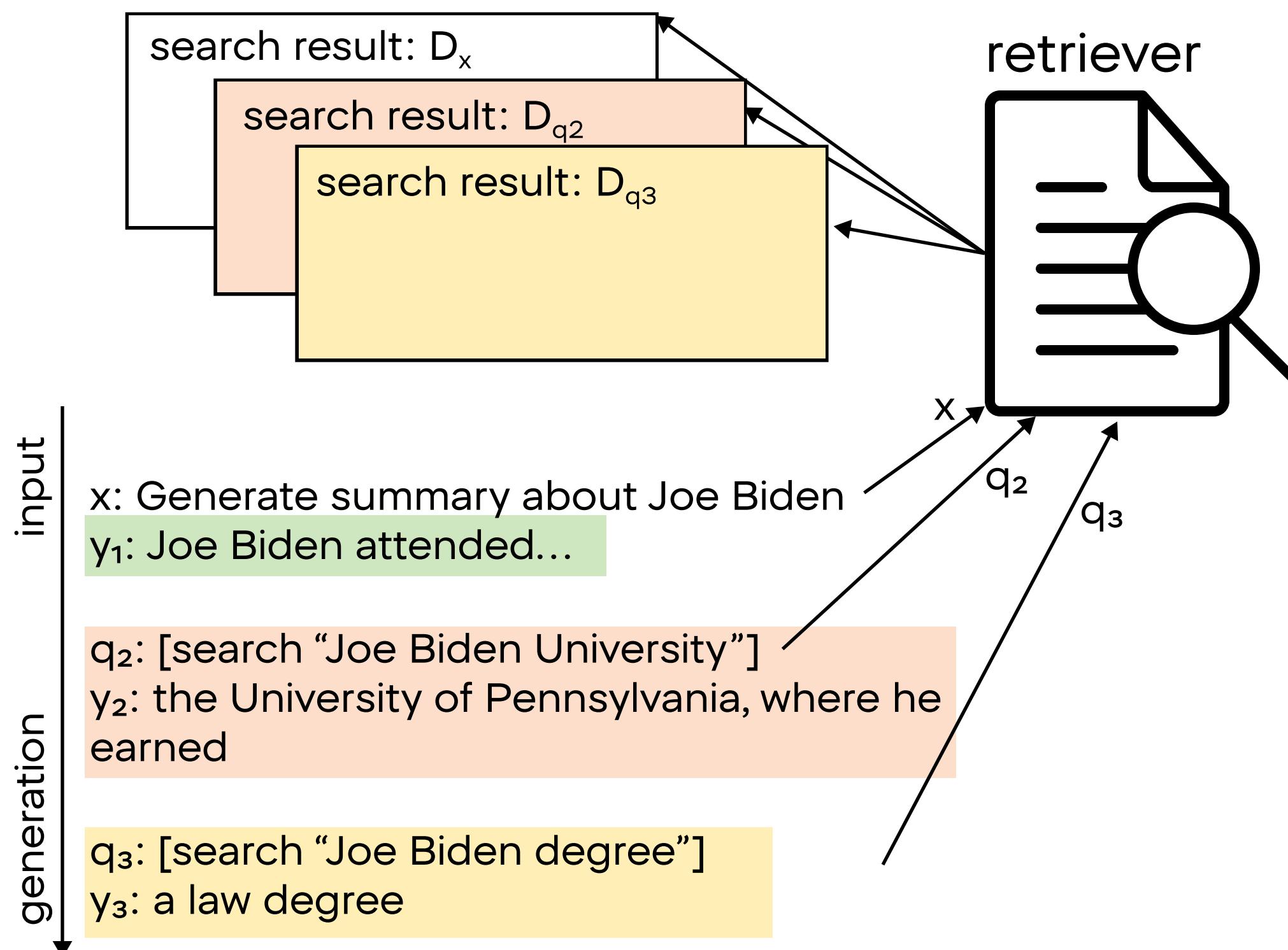


основная идея:

при помощи промптов вычисляем метаданные из запроса, сравниваем их с имеющимися в коллекции

FLARE

FORWARD-LOOKING ACTIVE RETRIEVAL

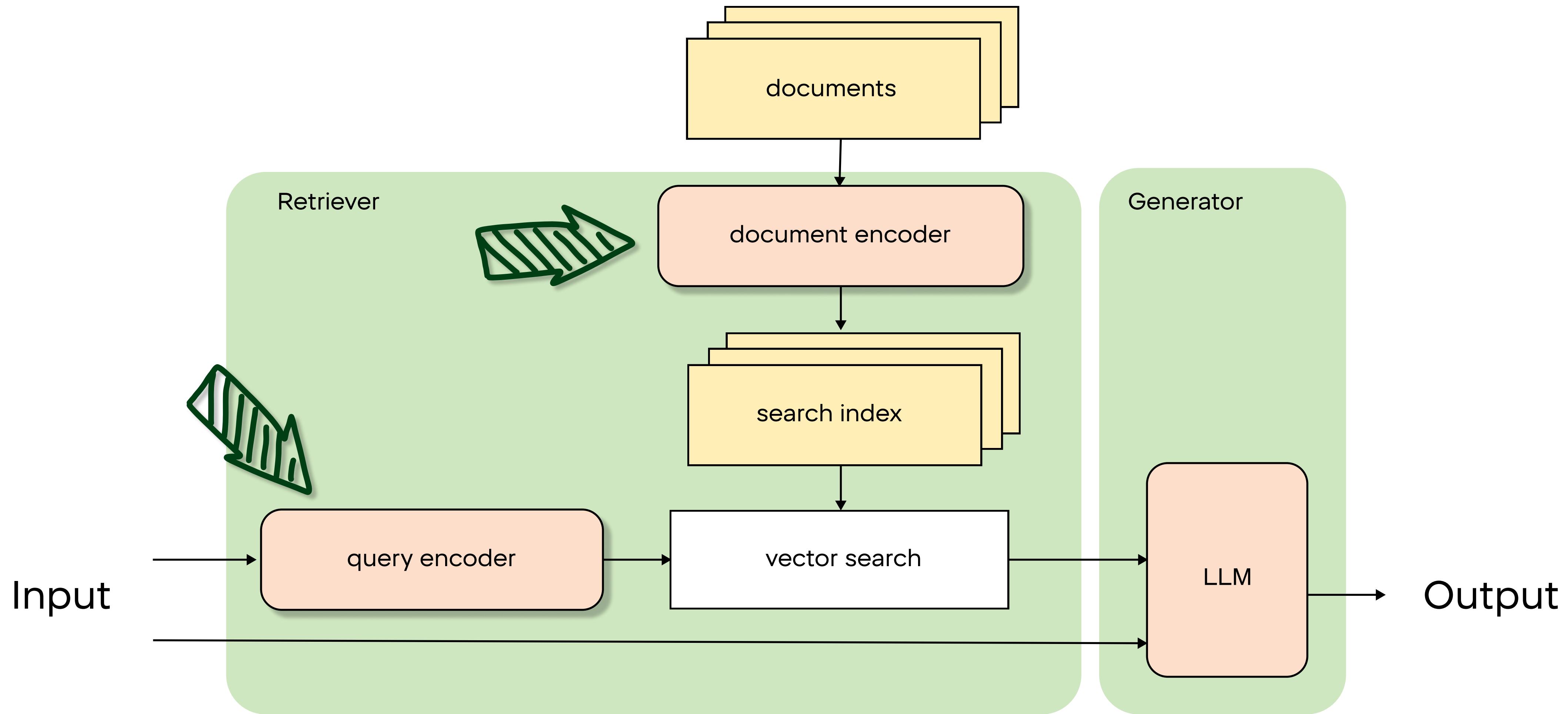


Идея:

- ☒ хотим повторять поиск по документам каждый раз, когда модель теряет уверенность
- ☒ теряет уверенность = генерирует токены низкой вероятности

SEARCH ENGINE

embeddings



embeddings

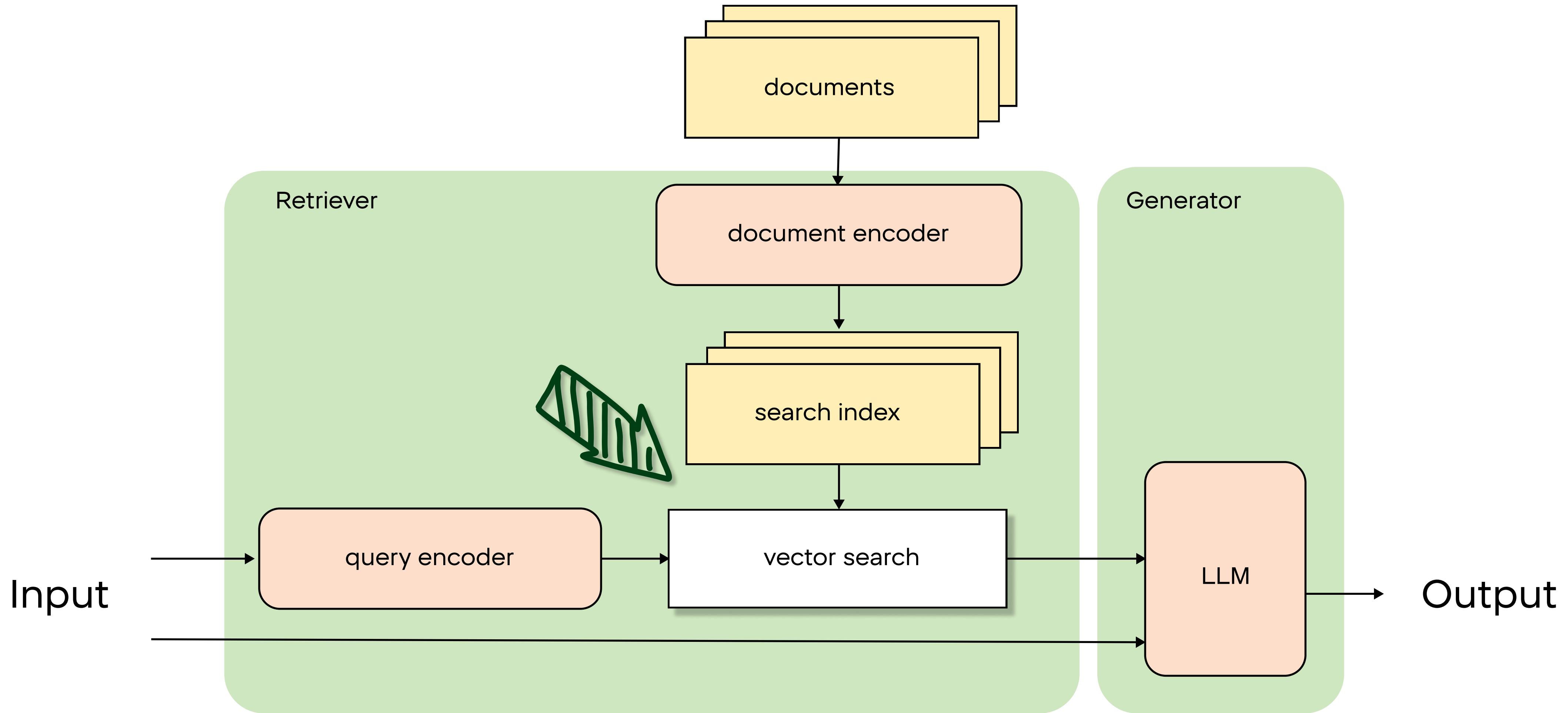
Что мы использовали?

- labse
- intfloat/multilingual-e5-large
- clips/mfaq

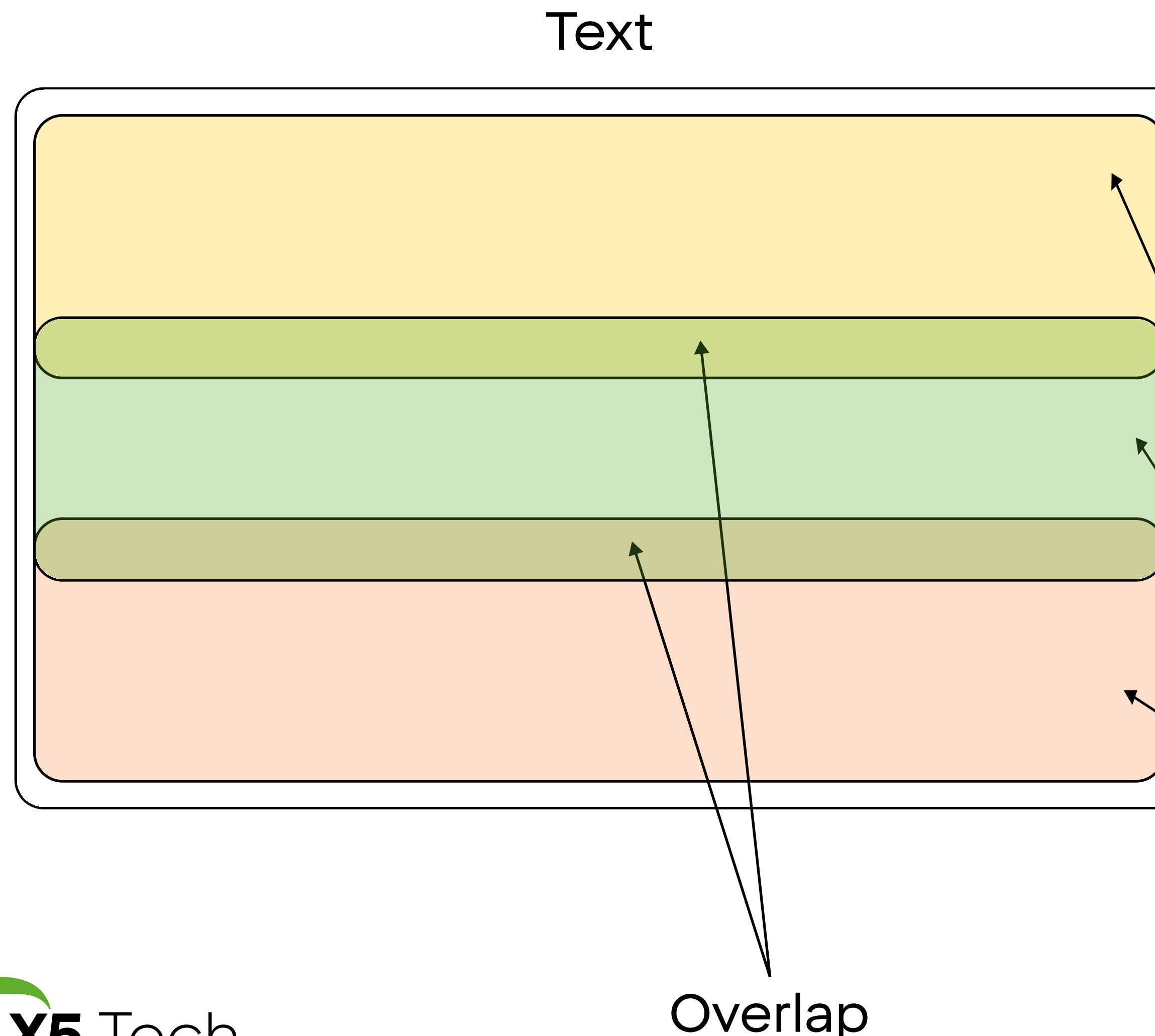
зачем пробовать разные?

качество эмбедингов непосредственно влияет на качество поиска похожих документов: чем лучше эмбединги работают с русским языком, тем лучше нужные документы отранжированы под конкретный вопрос

retrieve



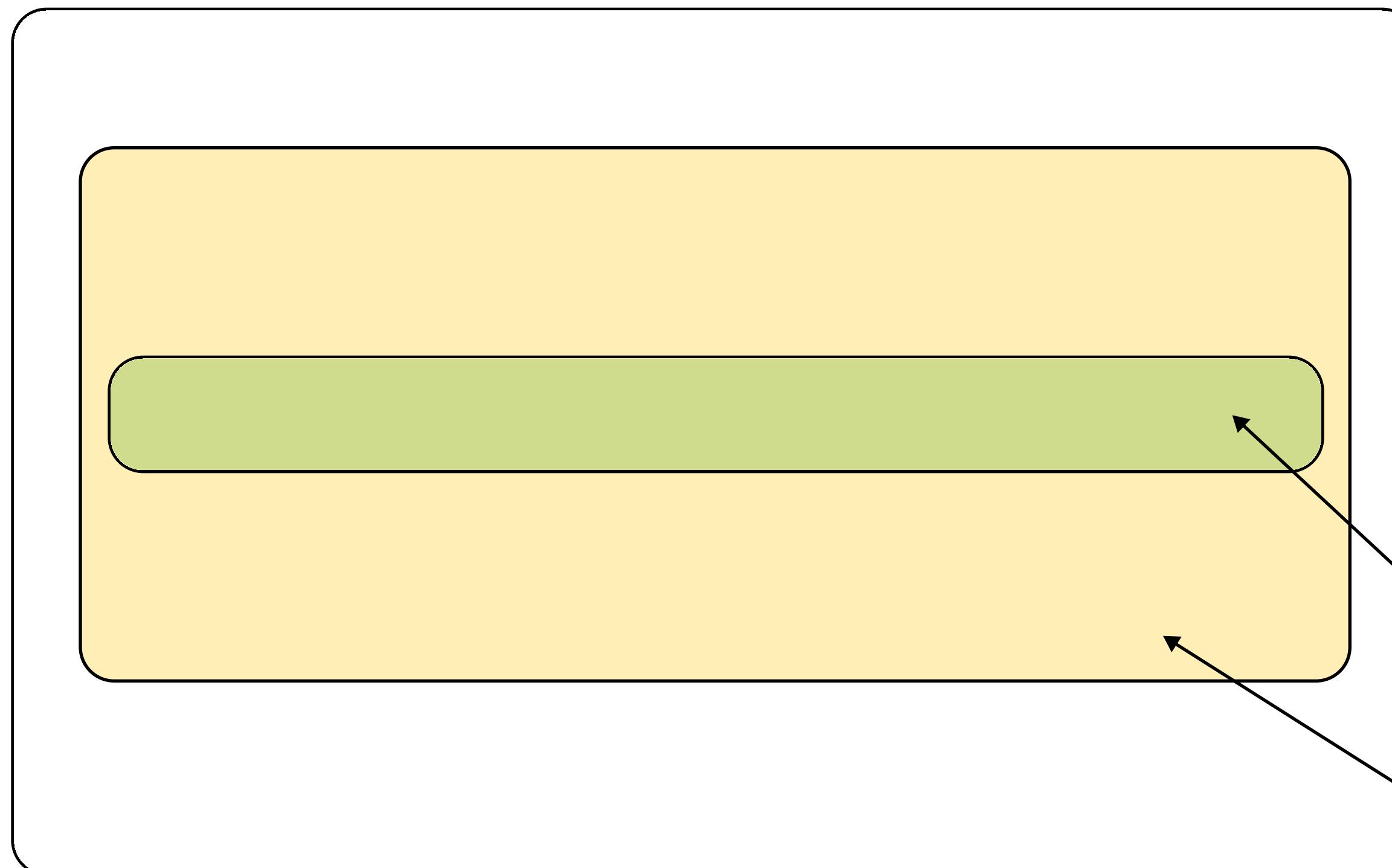
retrieve: chunking



- probuem разные размеры чанков, исходя из логики данных
- желательно, чтобы чанки отличались семантически
- помним про ограничение длины последовательности у би-энкодеров

retrieve: parent-document

Text



используем маленькие
чанки, каждый из которых
хранит ссылку на
родительский чанк большего
размера

Subchunk

Parent chunk

retrieve: ensemble

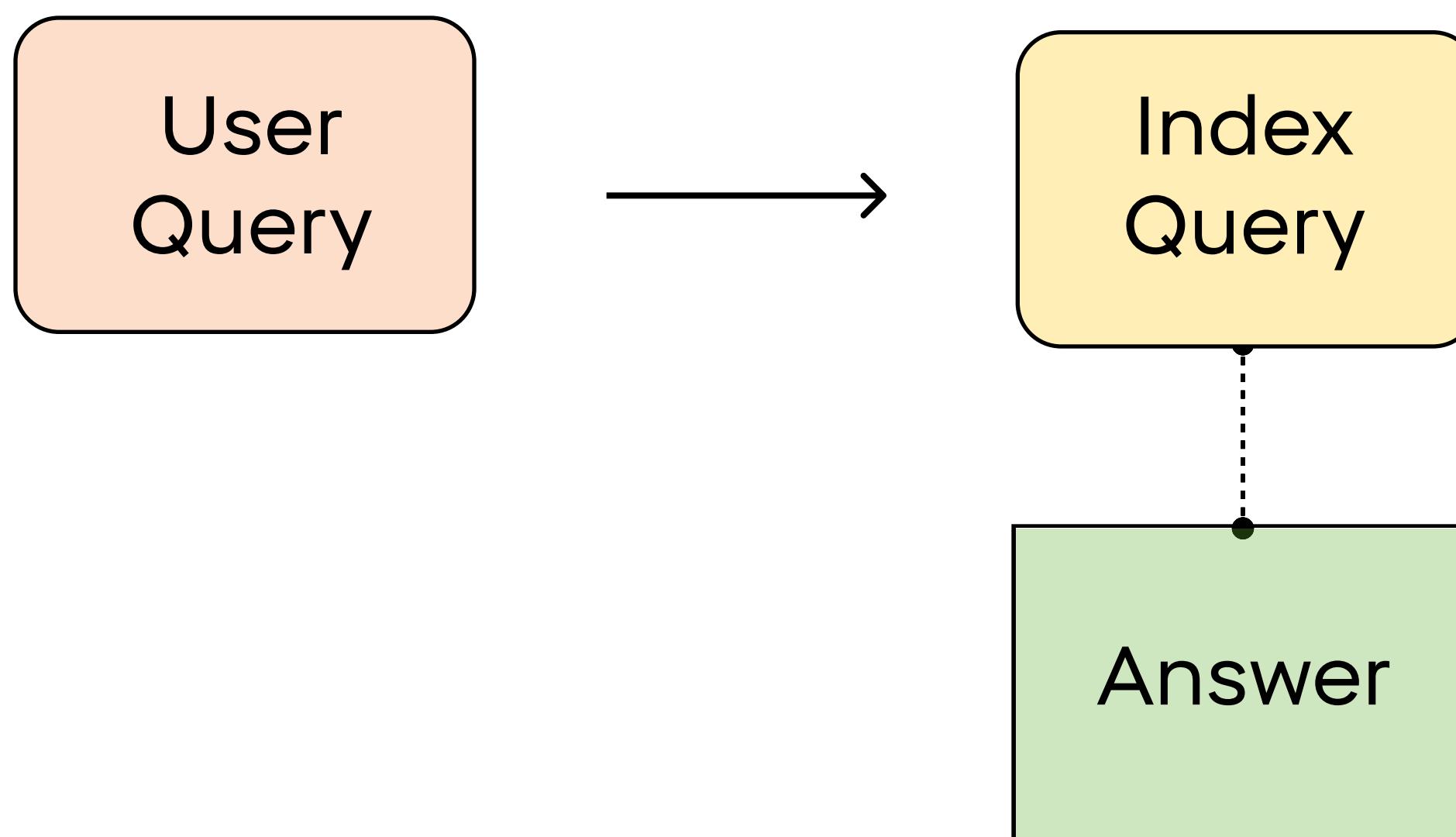
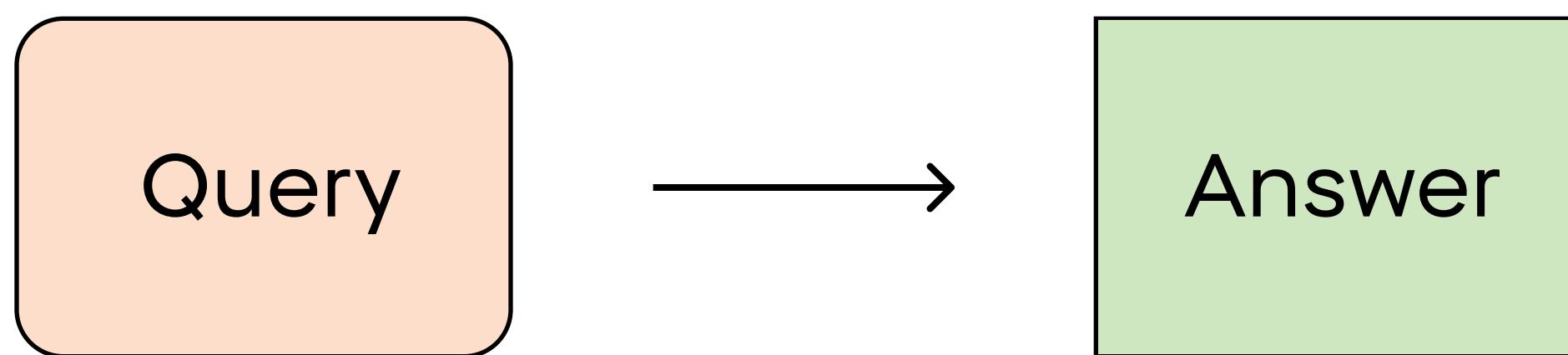
- 1. FAISS (baseline)
- 2. SVM
- 3. kNN
- 4. BM25
- 5. TFIDF

используют наши
эмбеддинги

используют свои
векторные
представления



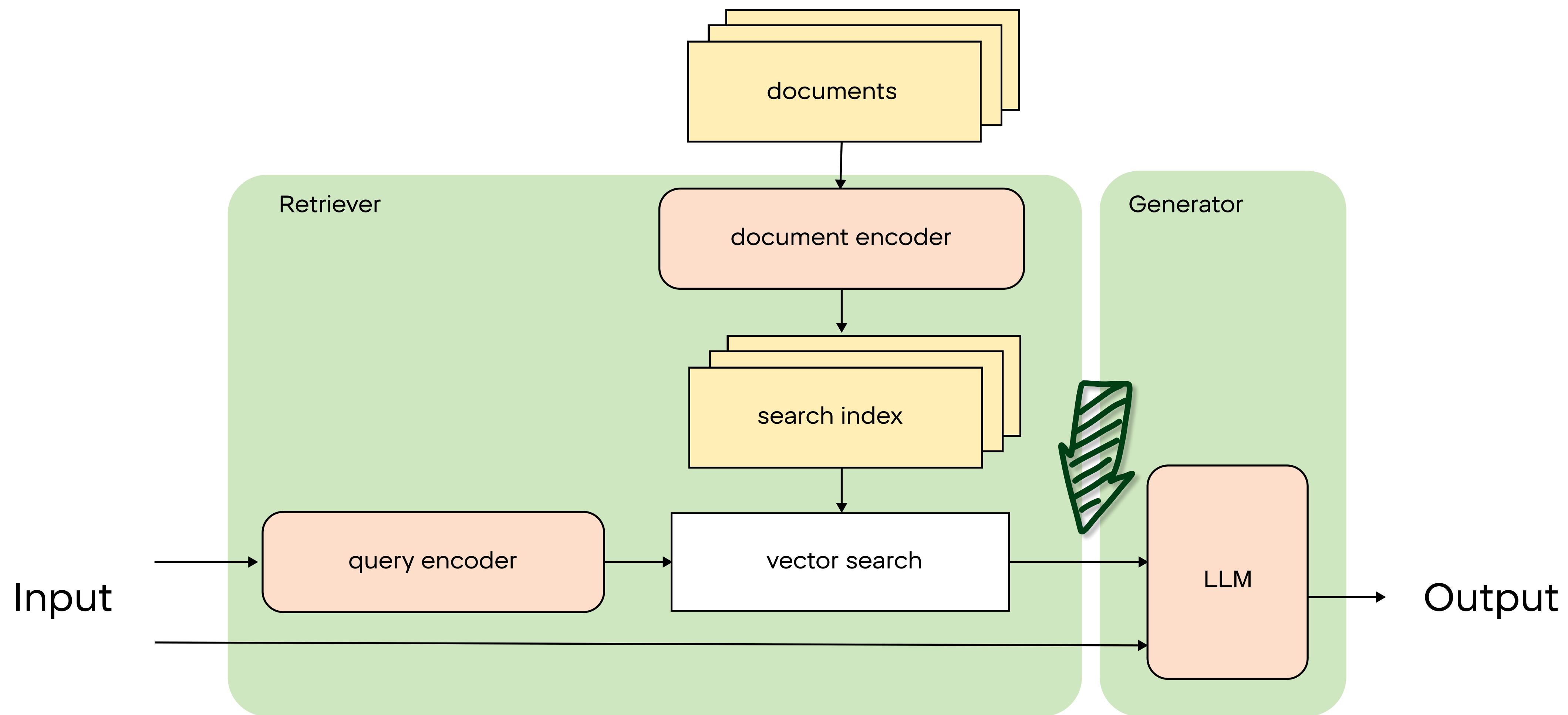
retrieve: symmetric search



- ☑ индексируем чанки вопросами и сравниваем запрос пользователя с индексом
- ☑ вопросы генерируем или собираем вручную
- ☑ со временем накапливаем статистику с реальными запросами и дополняем индекс

POSTPROCESSING

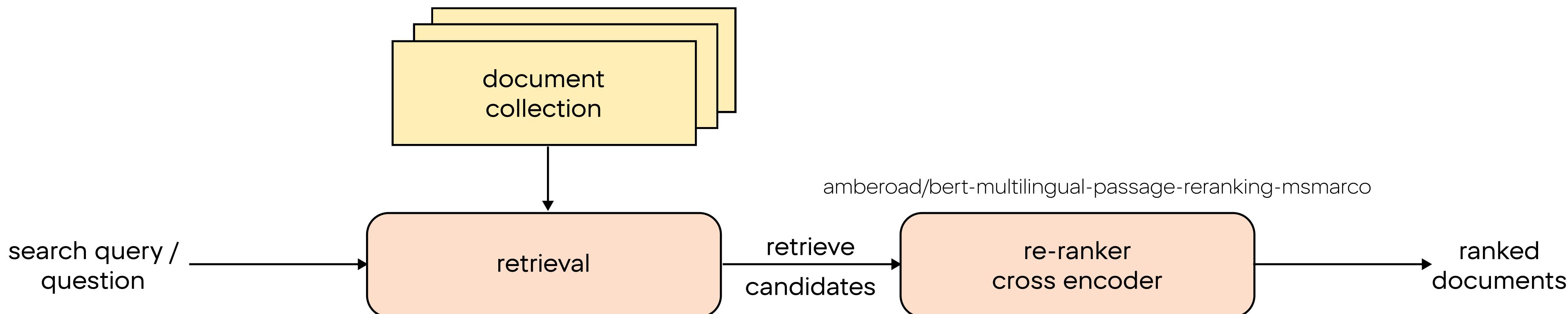
postprocessing



ReRank

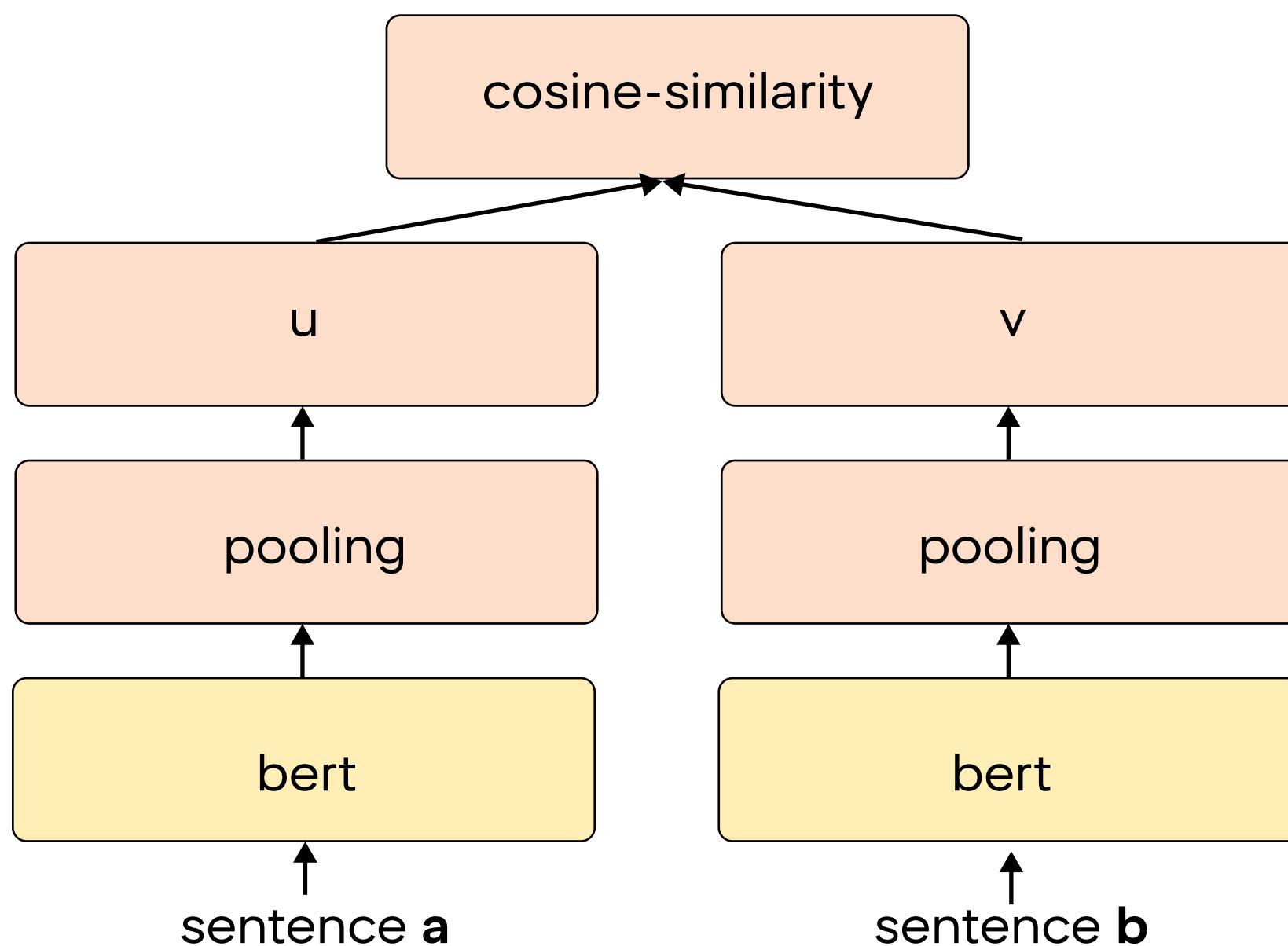
идея:

- Проиндексировать корпус чем-то легковесным для первичного отбора
- Вычислить их косинусное расстояние между парами запрос-ответ с помощью би-энкодера
- Взять топ К кандидатов и переранжировать их через кросс-энкодер

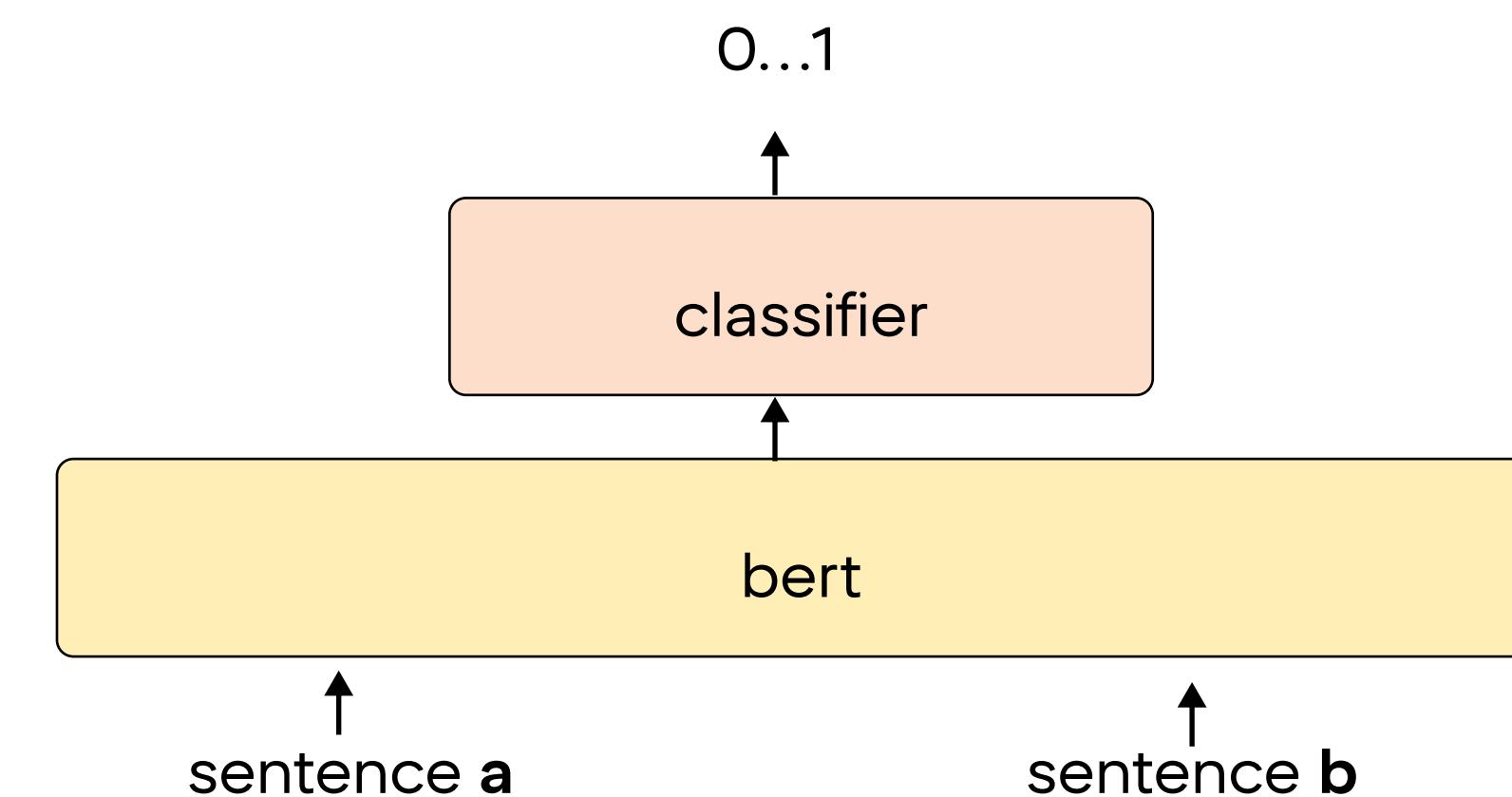


ReRank

bi-encoder



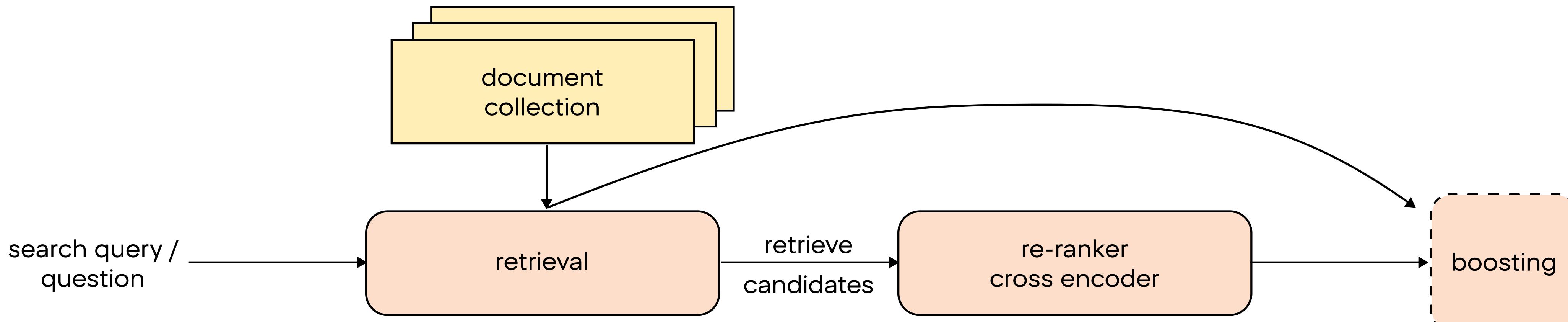
cross-encoder



ReRank++

идея:

Давайте засунем скоры в Boosting и переранжируем.



3

КАК ОТВЕЧАТЬ?

GENERATION

LLM prompting

пробуем:

- n-shot prompting
- cot (chain-of-thought)
- prompt chaining

помним:

- special tokens (bos, eos, etc.)
- prompt separation
- recency bias

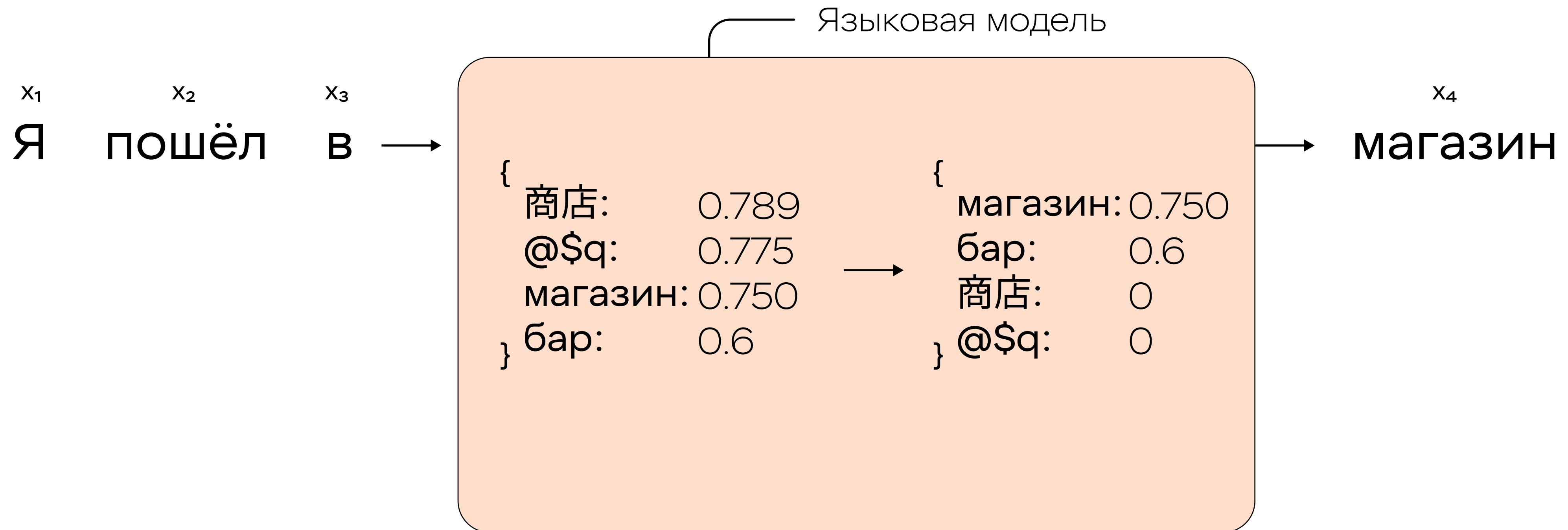
LLM generation parameters

- temperature
- top k
- top p
- repetition penalty
- presence penalty (could be negative)



logitsprocessor

Идея: зануляем вероятности для ненужных нам токенов



4

КАК ОЦЕНИТЬ РЕЗУЛЬТАТЫ?

метрики

- метрики расстояния между текстами
- метрики ранжирования
- llm-based метрики

метрики расстояния

- расстояние хэмминга
- расстояние левенштейна
- косинусное расстояние

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

в чём минусы?

не очень хорошо
реагирует на

грамматику

галлюцинации и
повторения

нужен эталон

метрики ранжирования

- map@k
- p@k
- ndcg@k

в чём минусы?

показывают только работу самого ретривера, а не всей rag системы в целом

нужен эталон

llm-based метрики

- groundedness – оценивает, насколько ответ модели обоснован с точки зрения контекста
- context relevance – оценивает, насколько контекст, добытый ретривером, соответствует вопросу пользователя

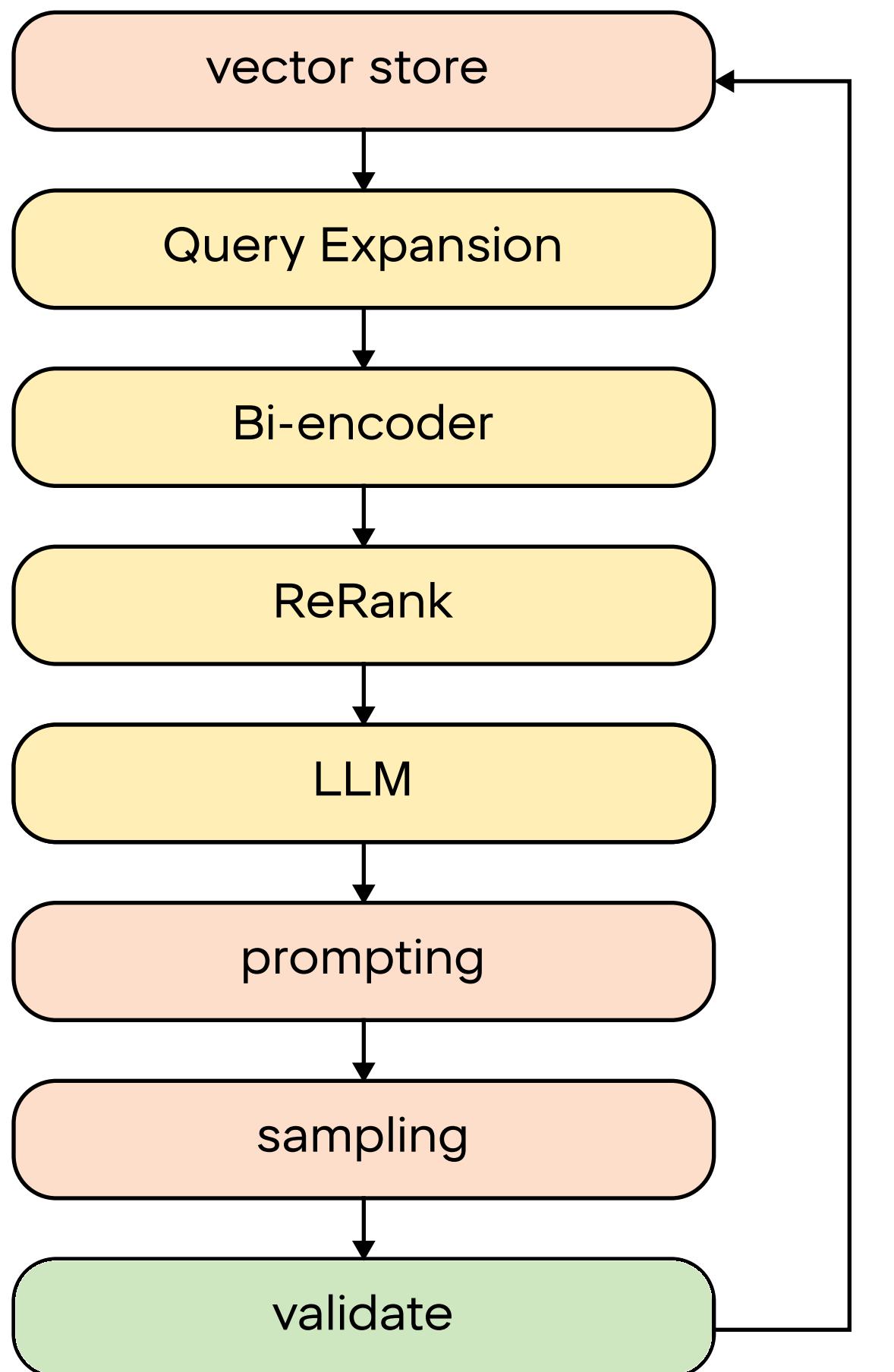
в чём минусы?

это все нестабильно, зависит от промптов, модели, погоды в казани и вообще субъективно и требует валидации руками

5

ЧТО В ИТОГЕ?

roadmap





ВАШИ ВОПРОСЫ

СПАСИБО ЗА ВНИМАНИЕ

МОЖЕТЕ СВЯЗАТЬСЯ С НАМИ В
TELEGRAM:

@AL_POTEKHIN

ОСТАВЬТЕ ВАШ ФИДБЭК ПО
НАШЕМУ ВЫСТУПЛЕНИЮ

