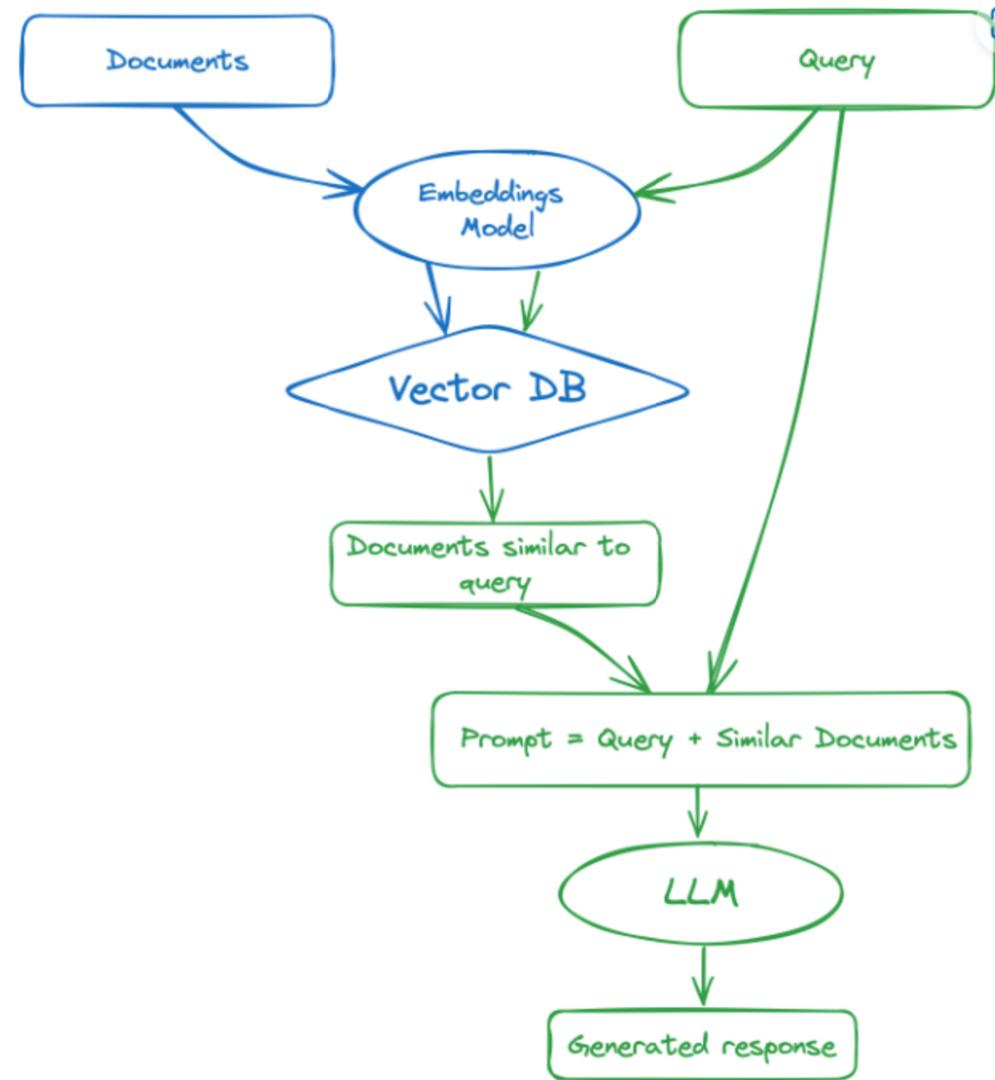




# **RAG:** от базового к продвинутому

Путилова Елена  
2 марта 2024

# RAG- workflow

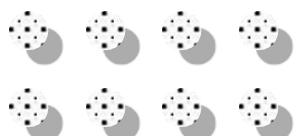


# Плюсы использования RAG

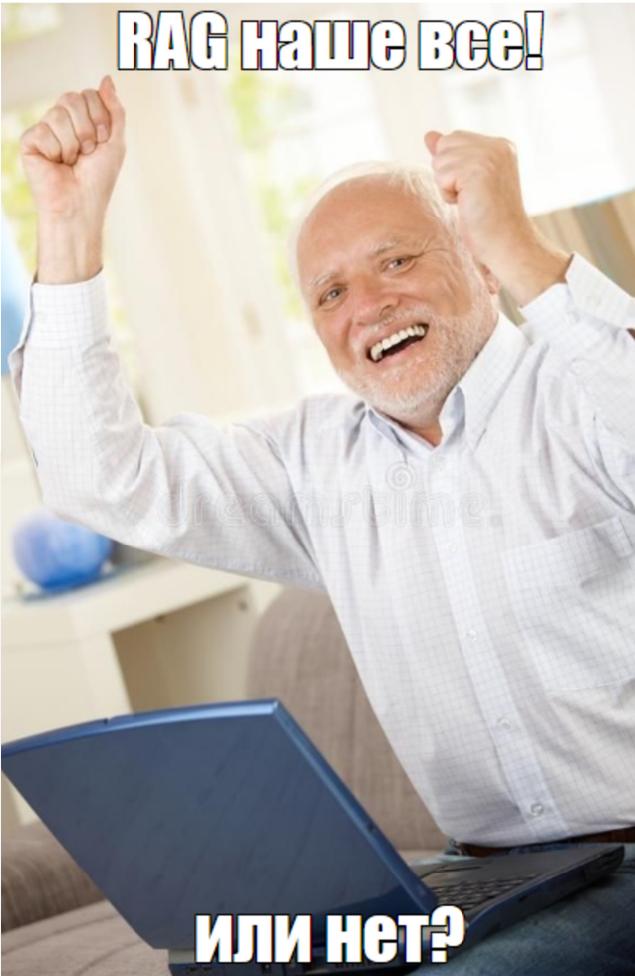
- Снижение галлюцинаций
- Альтернатива дорогому дообучению
- Модель быстро «узнает» специфический домен знаний
- Модель получает доступ к актуальной информации



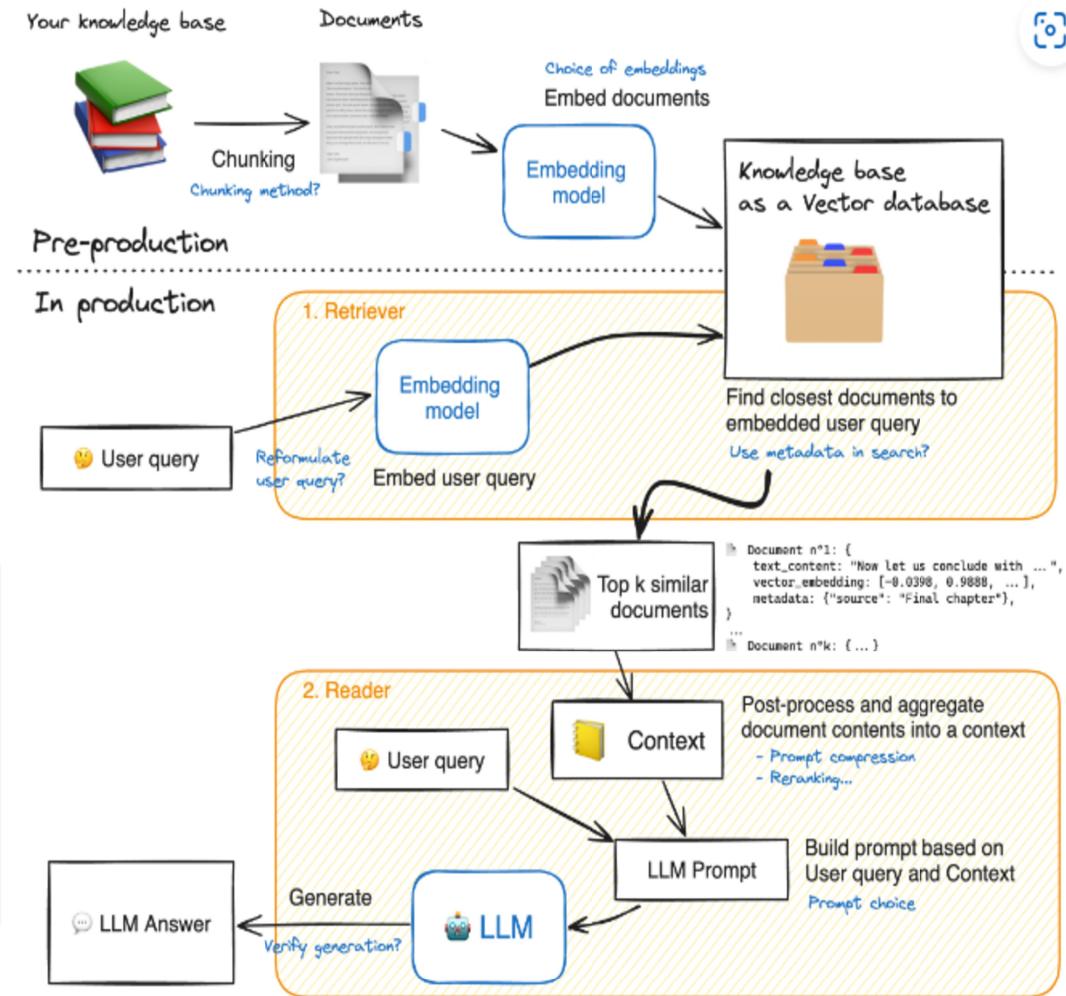
Illustrations by Pixeltrue on  
[icons8](#)



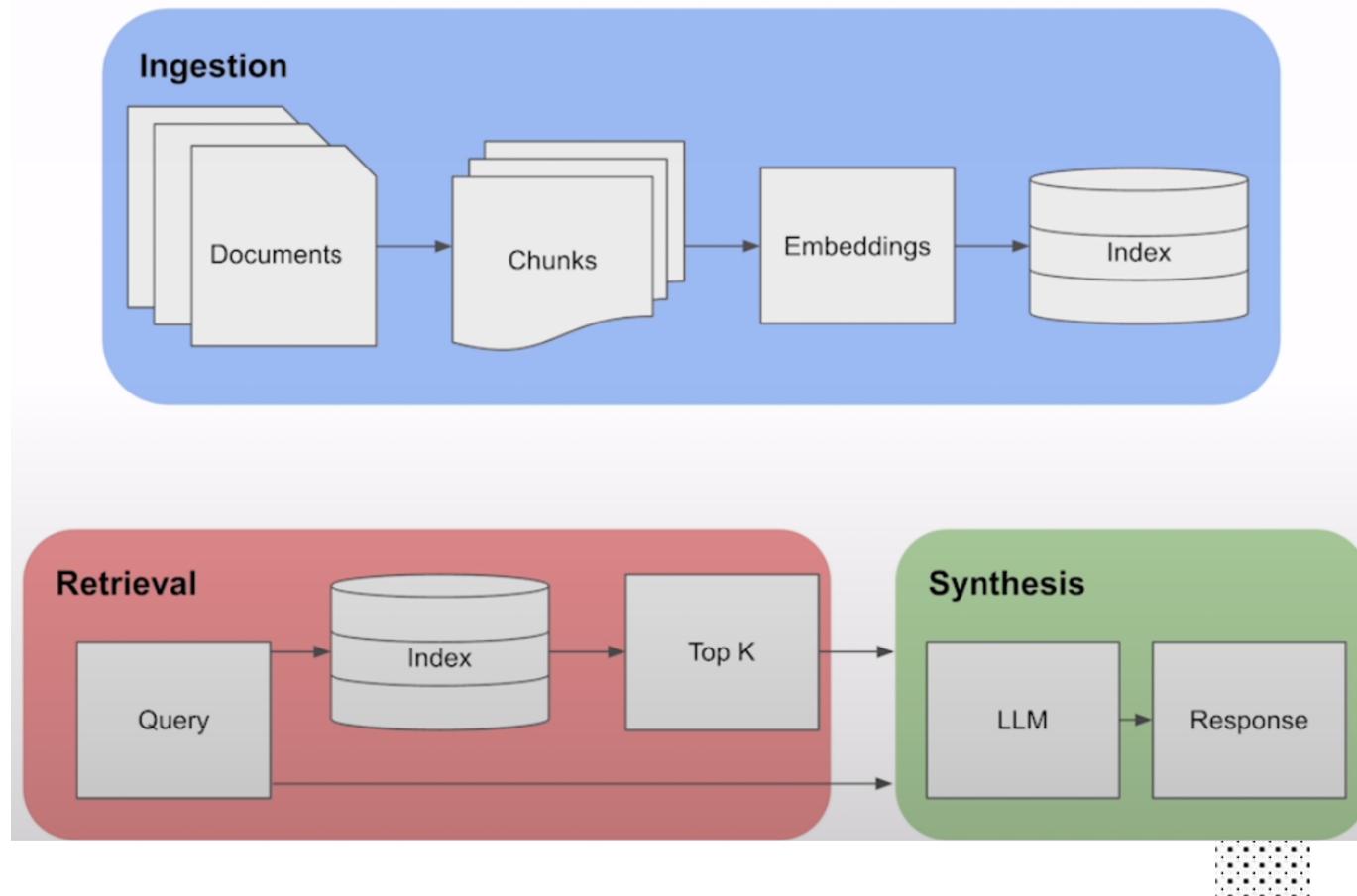
# Advanced RAG: зачем?



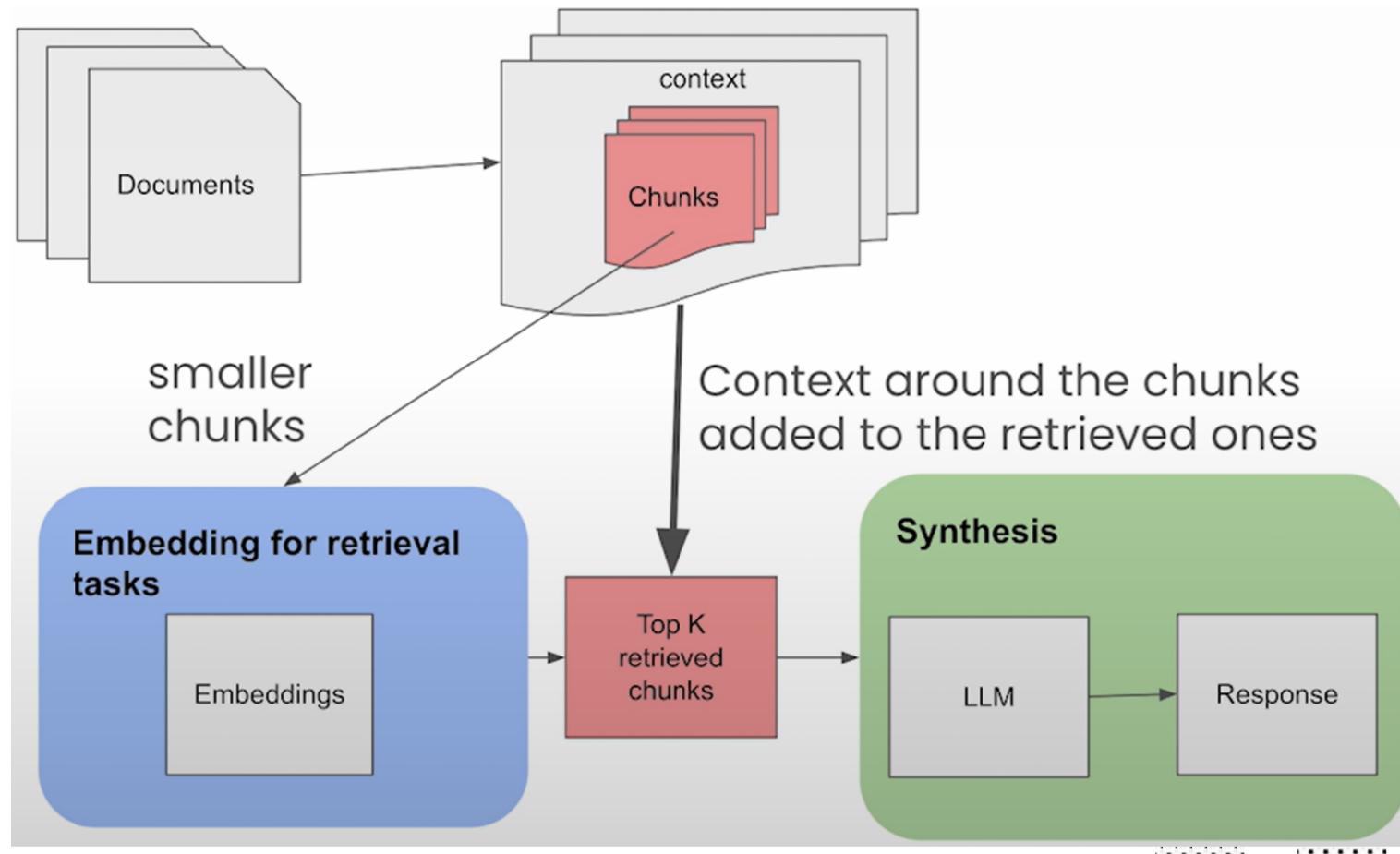
# Advanced RAG



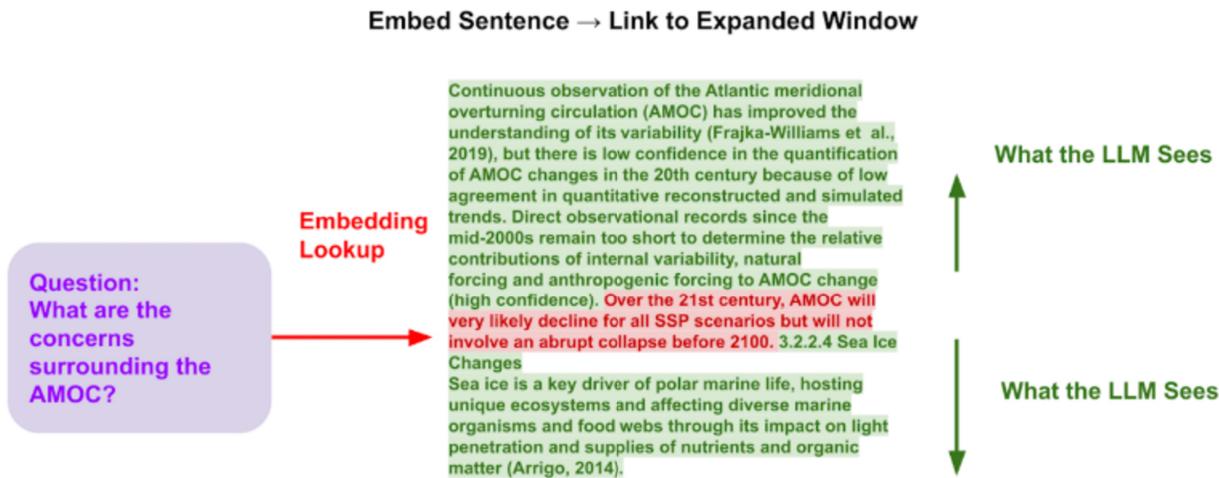
# Advanced RAG from LlamaIndex: what is problem?



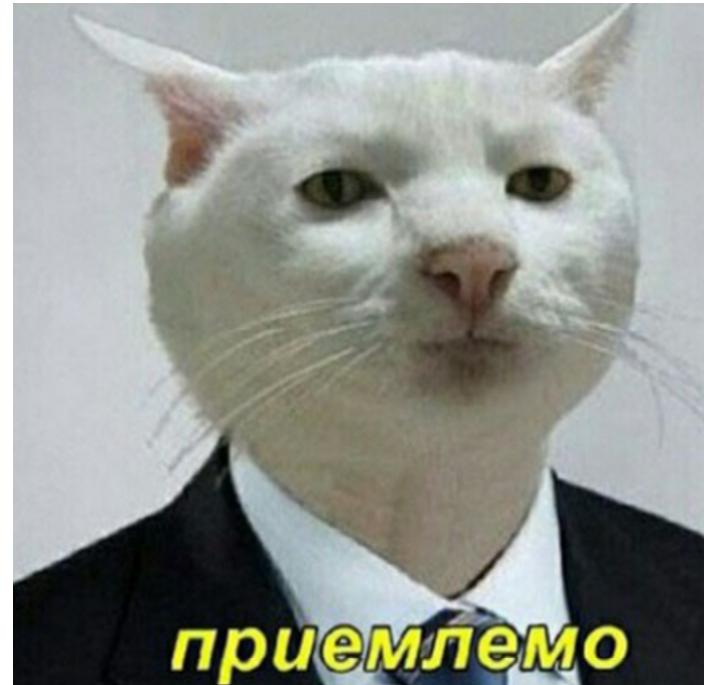
# Advanced RAG from LlamaIndex: sentence-window retrieval



# Advanced RAG from LlamaIndex: sentence-window retrieval



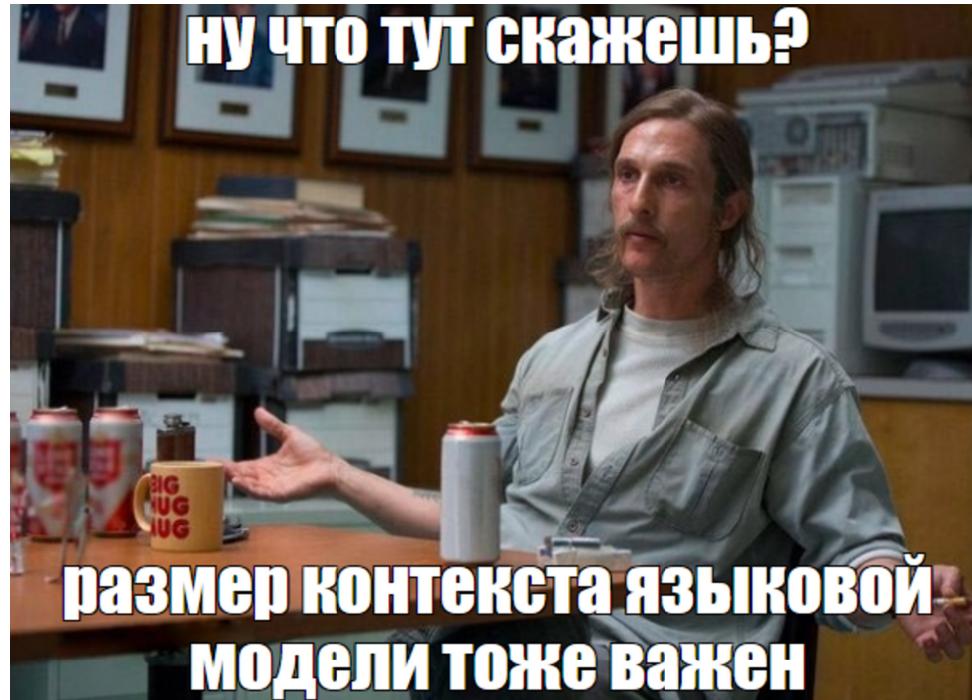
# Advanced RAG from LlamaIndex: результаты



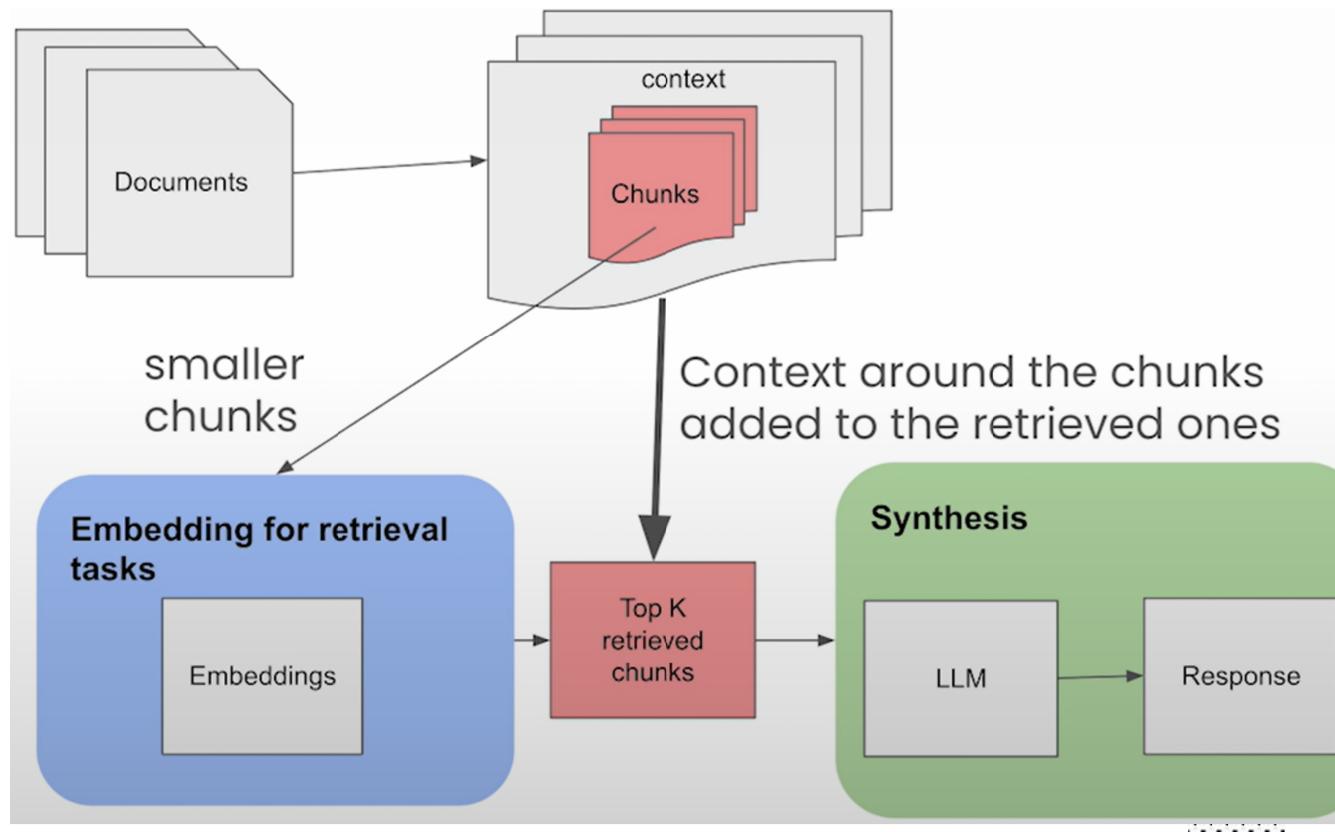
# **Advanced RAG: smart chunk's splitting from HuggingFace**

- размер chunk не должен превышать длину контекстного окна эмбеддинг-модели
- можно ввести дополнительную ранжировку

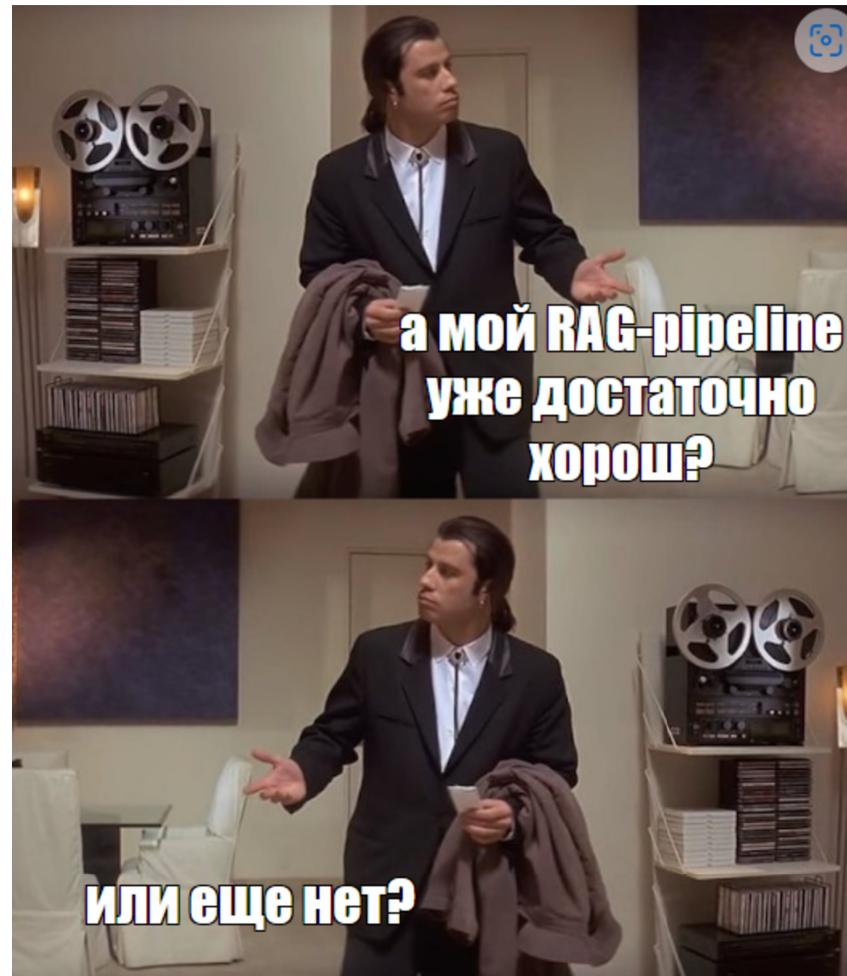
# Advanced RAG от HuggingFace: результаты



# Advanced RAG от HuggingFace: результаты



# RAG: как оценить



# RAG: ragas score

**generation**

**faithfulness**

**answer relevancy**

**retriever**

**context precision**

**context recall**



# RAG: как можно улучшить

- при разбивке на chunks ориентироваться на ограничение контекстного окна языковой модели и на размер смысловых частей в тексте (обычно это абзацы)
- использовать один из вариантов small2big подхода на pre-postretrieval стадиях
- подобрать актуальную эмбеддинг модель  
<https://huggingface.co/spaces/mteb/leaderboard>

## **Источники:**

- <https://huggingface.co/learn/cookbook>
- <https://python.langchain.com>
- <https://docs.llamaindex.ai/en/stable/>
- <https://www.deeplearning.ai/>
- <https://docs.ragas.io/>