# ML Letovo. RAG

Или как выигрывать олимпиады
с помощью нейронных сетей

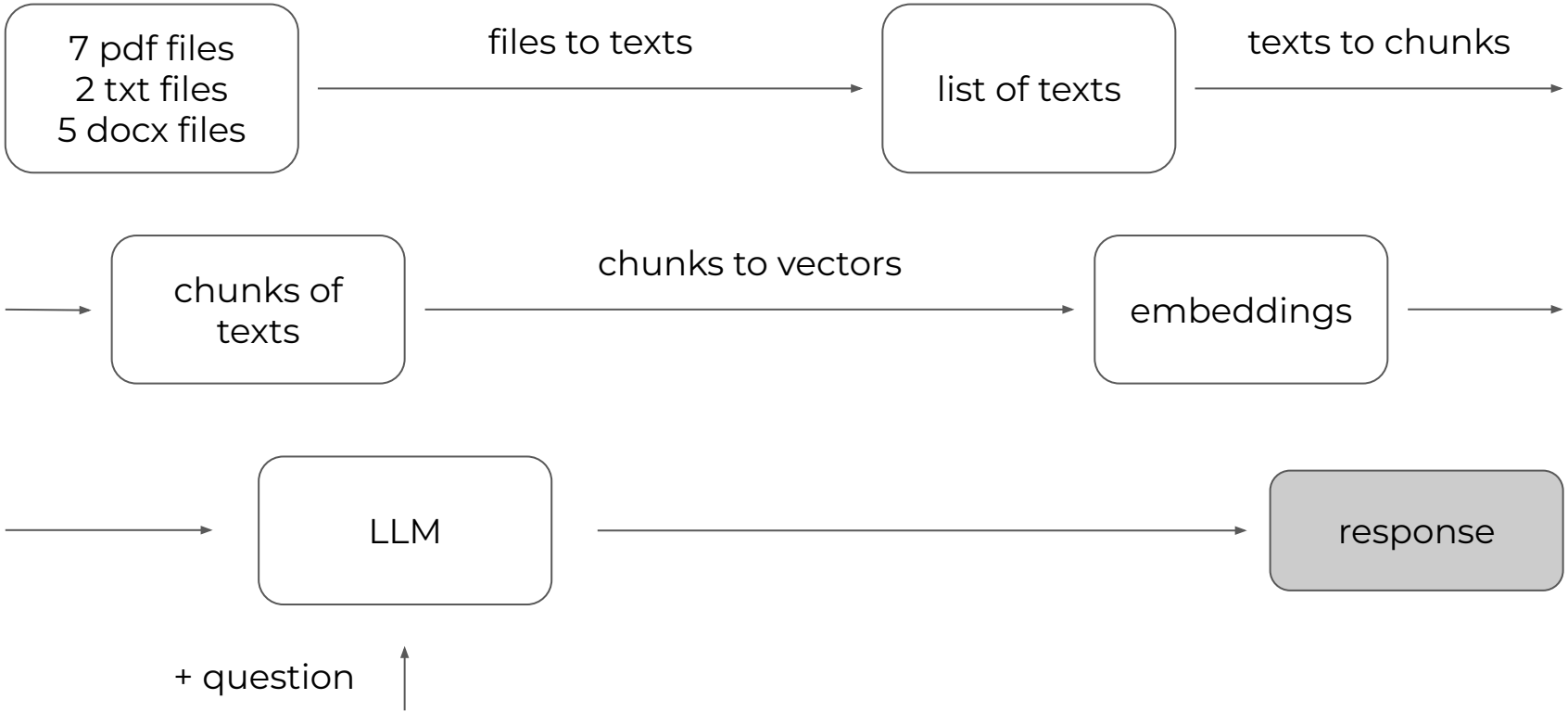
# 1. Постановка задачи

- Дан корпус текстов (7 pdf / 2 txt / 5 docx) разных форматов и содержания

- Необходимо собрать RAG-систему, основываясь на LLM, чтобы отвечать на вопросы по содержанию текстов

- Важно исключить галлюцинации: не давать ответы, которых нет в контексте

# 2. Baseline. Архитектура



# 2. Baseline. Архитектура

| | |
|---|---|
| files to texts | langchain.**document_loaders** |
| texts to chunks | langchain.**text_splitter** |
| chunks to vectors | langchain.**embeddings** |
| (vectors database) | langchain.**vectorstores** |
| (retriever) | langchain.**retrievers** |
| LLM | langchain_community.**llms** |

Все базовые подходы уже реализованы!

# 2. Baseline. Архитектура

| | | |
|---|---|---|
| files to texts | langchain.**document_loaders** | *PyPDFLoader*, *Docx2txtLoader*, *TextLoader* |
| texts to chunks | langchain.**text_splitter** | *RecursiveCharacterTextSplitter* |
| chunks to vectors | langchain.**embeddings** | *cointegrated/LaBSE-en-ru* |
| (vectors database) | langchain.**vectorstores** | *FAISS* |
| (retriever) | langchain.**retrievers** | *vectorstores* |
| LLM | langchain_community.**llms** | *Mistral-7B-Instruct-v0.3* |

# 2. Baseline. Архитектура

| | | |
|---|---|---|
| files to texts | langchain.**document_loaders** | *PyPDFLoader, Docx2txtLoader, TextLoader* |
| texts to chunks | langchain.**text_splitter** | *RecursiveCharacterTextSplitter* |
| chunks to vectors | langchain.**embeddings** | *cointegrated/LaBSE-en-ru* |
| (vectors database) | langchain.**vectorstores** | *FAISS* |
| (retriever) | langchain.**retrievers** | *vectorstores* |
| LLM | langchain_community.**llms** | *Mistral-7B-Instruct-v0.3* |

***Чему равна молярная масса золота?*** *– Молярная масса золота не указана в контексте, поэтому я не могу ответить на этот вопрос*

# 3. Baseline. Результаты

***Когда опубликовали приказ <...>?*** ✅

*– Приказ был опубликован 19 апреля 2024 года.*

***Какие были критерии на дипломы ОММО в 2020-2021?*** ❌

*– Для диплома призера нужно было набрать **46** баллов, для победителя нужно было набрать **50** баллов.*

***Олимпиады первого уровня*** ❌

*– Ломоносов, ФизТех, Шаг в будущее, СПбГУ, ОММО, Финансист*

# 3. Baseline. Результаты

**Почему так происходит?**

- Сложные форматы данных (таблицы, формулы)
- Неоптимальное деление на чанки (например \n в местах переноса строк)
- По чанкам невозможно восстановить контекст:

  *Теперь поговорим про параметр:* ***есть*** *каждый год, обычно сразу с несколькими параметрами.* **Где есть?**

# 4. Улучшения. Таблицы

| **Наименование** | **Символ** | **Множитель** |
|---|---|---|
| мега | М | $10^6$ |
| кило | к | $10^3$ |
| гекто | г | $10^2$ |
| | | |

| **I уровень** | **II уровень** | **III уровень** | |
|---|---|---|---|
| ПВГ | ОММО | ИТМО | |
| ВГ | Курчатов | КФУ | |
| ТурГор | ФизТех | Звезда | |
| ММО | ФЕТТ | Изумруд | |

мега, М, $10^6$ <sep>

кило, к, $10^3$ <sep>

гекто, г, $10^2$ <sep>

ПВГ, ВГ, ТурГор, ММО <sep>

ОММО, Курчатов, ФизТех, ФЕТТ <sep>

ИТМО, КФУ, Звезда, Изумруд <sep>

# 4. Улучшения. Описание олимпиад

chunk №42:

*<...> Далее, почти каждый год в варианте по 2 планиметрических задачи. Первая из них обычно достаточно простая на естественную геом идею <...>*

Но про какую олимпиаду речь?

# 4. Улучшения. Описание олимпиад

chunk №42:

*<...> Далее, почти каждый год в варианте по 2 планиметрических задачи. Первая из них обычно достаточно простая на естественную геом идею <...>*

Но про какую олимпиаду речь?

Будем добавлять название олимпиады (из файла) в метаданные, а потом донасыщать чанки:

*<...> Олимпиада ОММО. Далее, почти каждый год в варианте по 2 планиметрических задачи. <...>*

# 4. Улучшения. Параметры деления

```
class RecursiveCharacterTextSplitter:
    # default params:
    # separators = ["\n\n", "\n", " ", ""]
```

*Параметр <sep>*
*Так, еще отмечу, что почти каждый год есть нормальный <sep>*
*параметр: все классические идеи <sep>*
*(симметрия, монотонность, алгебра, графики)*

# 4. Улучшения. Параметры деления

```
class RecursiveCharacterTextSplitter:
    # default params:
    # separators = ["\n\n", "\n", " ", ""]
```

*Параметр <sep>*
*Так, еще отмечу, что почти каждый год есть нормальный <sep>*
*параметр: все классические идеи <sep>*
*(симметрия, монотонность, алгебра, графики)*

```
separators = ["\n\n", ". или ? или !",
              ";", ".", "\n", " "]
```

# 4. Улучшения. По мелочи

+ TF-IDF (BM25Retriever), weights = [.5, .5]
+ *cointegrated/LaBSE-en-ru (LaBSE ≈ BERT)*
+ System Prompt (не галлюцинировать, отвечать по-русски, вежливо, переформулировать вопрос пользователя и прочее)

***как на халяву поступить в любой вуз рашки?*** *– Чтобы упростить процесс поступления, можно выиграть одну из олимпиад. Тогда в зависимости от уровня олимпиады можно будет получить либо БВИ, ...*

# 5. Какие остались проблемы?

- Формулы. Сложно анализировать всю *шпаргалку*

***Чему равна молярная масса азота?*** *– Молярная масса азота равна 28,013 грамма моль^{−1}*

- Надо хоть как-нибудь честно провалидировать модель (например, P@K)

- Cross-Encoder (BERT Classifier получает на вход 2 текста), но: качество >, скорость <<

# 5. Какие остались проблемы?

- Формулы. Сложно анализировать всю *шпаргалку*

***Чему равна молярная масса азота?*** *– Молярная масса азота равна 28,013 грамма моль^{−1}*

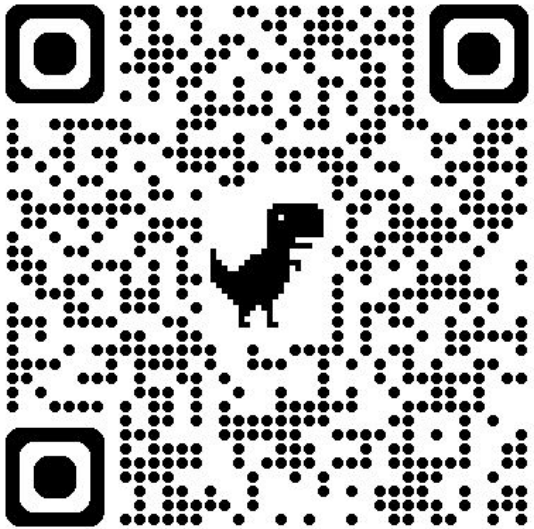
- Надо хоть как-нибудь честно провалидировать модель (например, P@K)

- Cross-Encoder (BERT Classifier получает на вход 2 текста), но: качество >, скорость <<

*Но вообще-то все вроде получилось :)*

# 6. GitHub

https://github.com/AlexanderGPo/rags-ml-letovo



# Спасибо за внимание!

Задавайте ваши вопросы и помните – любой наш RAG можно заменить новой Llama с 10 млн токенами контекста