



Проектный курс по Большим Языковым
Моделям

ВШЭ

Машинное обучение и
высоконагруженные системы

Реализация RAG системы на основе статей с arxiv.org

[GitHub](#)

Кошелев Никита
Югай Данила
2025



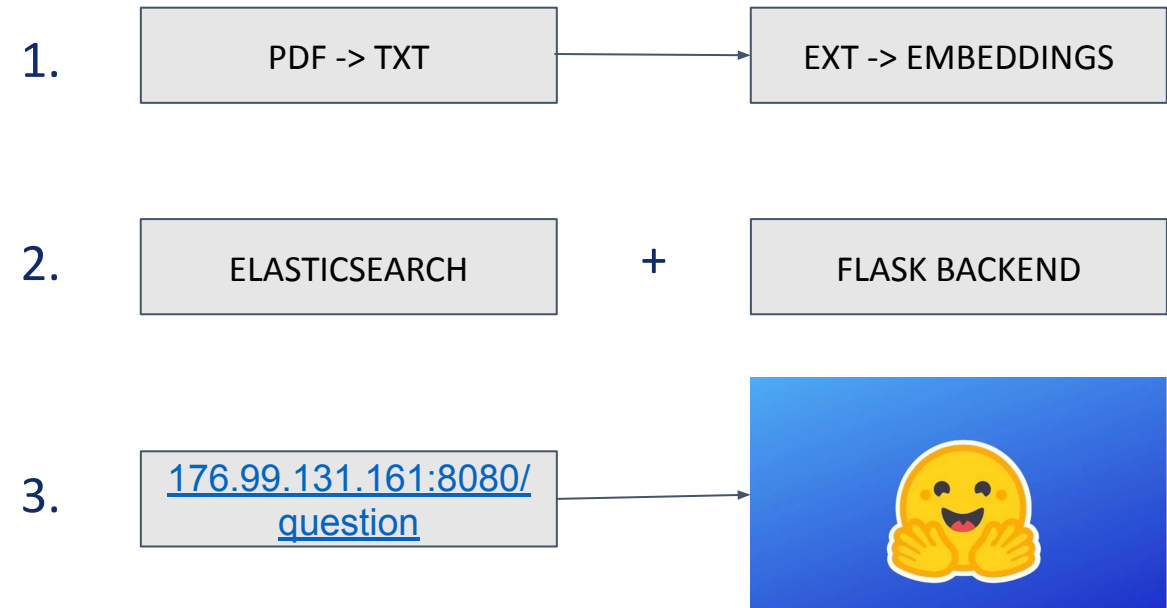
Содержание

1. Helicopter view
2. От PDF до Эмбедингов на текст
3. Elasticsearch и сервис
4. PHI 3.5 и деплой сервиса
5. Метрика и примеры работы
6. Дальнейшее развитие системы



Helicopter view

- 1) Обработка документов:
Перевод в текстовый формат, разбиение на чанки,
создание эмбедингов
- 2) Поднятие сервиса:
2 AI модели + ElasticSearch, ее заполнение.
- 3) Интеграция модели в Hugging Face Spaces



Модели, используемые в системе: [BGE-3](#), [Qwen2.5-VL-3B-Instruct](#), [Phi-3.5-mini-instruct](#)



От PDF до Эмбедингов на текст

- 1) **Перевод изображений в текст:**
 - a) Подбор промта - хотим переводить таблицы в текст
 - b) Сложности с переводом графиков в текст - кажется нужна модель побольше
- 2) **Перевод изображений в текст:**
 - a) Разбиение текста на чанки
 - b) Каждому чанку - эмбединг

VAST.AI, GPU: 4070ti 16GB

VL model: [Qwen2.5-VL-3B-Instruct](#) 3.75B params

Sentence-similarity: [BGE-3](#) 3.8B params

Оригинальная статья (PDF)

| Training Costs | Pre-Training | Context Extension | Post-Training | Total |
|-------------------|--------------|-------------------|---------------|----------|
| in H800 GPU Hours | 2664K | 119K | 5K | 2788K |
| in USD | \$5.328M | \$0.238M | \$0.01M | \$5.576M |

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.

and generation length.

We evaluate DeepSeek-V3 on a comprehensive array of benchmarks. Despite its economical training costs, comprehensive evaluations reveal that DeepSeek-V3-Base has emerged as the strongest open-source base model currently available, especially in code and math. Its chat version also outperforms other open-source models and achieves performance comparable to leading closed-source models, including GPT-4o and Claude-3.5-Sonnet, on a series of standard and open-ended benchmarks.

Статья в текстовом формате

Table 1 | Training Costs of DeepSeek-V3

| Training Costs | Pre-Training | Context Extension | Post-Training | Total |
|-------------------|--------------|-------------------|---------------|----------|
| in H800 GPU Hours | 2664K | 119K | 5K | 2788K |
| in USD | \$5.328M | \$0.238M | \$0.01M | \$5.576M |

Table 1 | Training Costs of DeepSeek-V3, Assuming the Rental Price of H800 is \$2 per GPU Hour.

We evaluate DeepSeek-V3 on a comprehensive array of benchmarks. Despite its economical training costs, comprehensive evaluations reveal that DeepSeek-V3-Base has emerged as the strongest open-source base model currently available, especially in code and math. Its chat version also outperforms other open-source models and achieves performance comparable to leading closed-source models, including GPT-4o and Claude-3.5-Sonnet, on a series of standard and open-ended benchmarks.

Эмбединг

```
array([[-0.04214 , -0.01968 , -0.03366 , ..., -0.04382 ,  0.04047 ,  
        0.00379 ]],
```



ElasticSearch и сервис



```
"properties": {  
  "text": {"type": "text"},  
  "title": {"type": "text"},  
  "embedding": {  
    "type": "dense_vector",  
    "dims": DIMS,  
    "index": True,  
    "similarity": "cosine",  
  },  
}
```



Flask

POST <HOST>/question

Request:

```
{  
  "question": str, - Сам вопрос  
  "use_context": bool - использовать ли контекст  
}
```

Response:

```
{  
  "answer": str, - Сам вопрос  
  "context": str - текст полученный из ElasticSearch  
  "question": bool - заданный вопрос,  
  "used_papers": list - названия статей, откуда взяли  
  данные  
}
```



Phi-3.5 и деплой сервиса

Модель обучена в марте 2023 и имеет ряд преимуществ:

- Легковесная, state-of-the-art модель своего поколения (2023 год)
- Контекст длины 128k токенов
- Мультиязычность

Деплой модели:

- One-Click Deployment - Hugging Face Spaces

immers.cloud, GPU: T4 16GB

LLM: [Phi-3.5-mini-instruct](#) 3.82B params

The screenshot shows a web interface for a RAG system deployed on Hugging Face Spaces. The browser address bar shows 'huggingface.co'. The page title is 'rag-system' by user 'Dangennadevich'. Below the title, there are 'like' and 'Running' indicators. The main section is titled 'Вопрос-ответ через ваш API'. It contains a form with a text input for 'Ваш вопрос' (Your question) with the placeholder text 'Сколько параметров у DeepSeek v3?'. Below the input is a dropdown menu for 'Использовать RAG?' (Use RAG?) set to 'true'. A button labeled 'Отправить запрос' (Send request) is below the form. The response section, titled 'Ответ системы' (System answer), displays the text: 'DeepSeek-V3, как указано в предоставленном тексте, имеет в общей сложности 671 миллиард параметров. Этот размер параметров отражает мощность модели и объемы обучения, которые она может обрабатывать.' Below the response is a 'Статус контекста' (Context status) section showing 'Контекст использован' (Context used) with a green checkmark. At the bottom, there is a footer that says 'Создано с помощью Gradio' (Created with Gradio) and a 'Настройки' (Settings) link.



Метрика и примеры работы

В качестве метрики посчитали Accuracy:

Модель с RAG: 0.86

Модель без RAG: 0.05

Ошибки модели с RAG:

- Неверно переводится графическая информация в текст

Ошибки модели без RAG:

- Модель совершенно не знает тематики и сама просит предоставить ей контекст
- Поверхностные суждения без погружения в тему
- Не может исправить опечатки и начинает галлюцинировать

Вопрос-ответ через ваш API

Ваш вопрос

What are the key differences between DeepSeek-R1-Zero and DeepSeek-R1 in terms of training methodology and performance?

Использовать RAG?

true

Отправить запрос

Ответ системы

DeepSeek-R1-Zero and DeepSeek-R1 differ primarily in their training methodologies and the resulting performance on reasoning tasks:

1. Training Methodology:
 - DeepSeek-R1-Zero: This model is trained exclusively through large-scale reinforcement learning (RL) without any supervised fine-tuning (SFT). It relies solely on the RL process to develop reasoning capabilities, starting from a base model (DeepSeek-V3-Base).

Вопрос-ответ через ваш API

Ваш вопрос

What are the key differences between DeepSeek-R1-Zero and DeepSeek-R1 in terms of training methodology and performance?

Использовать RAG?

false

Отправить запрос

Ответ системы

1. **Initialization**:
 - **DeepSeek-R1**: The model parameters are initialized using a standard method, which could be **random initialization**, Xavier/Glorot, He initialization, or another method suitable for the activation functions used in the network.



Метрика и примеры работы

Ваш вопрос

What method was used for the Reward Model in deepseek m3?

Использовать RAG?

false

The DeepSeek M3 (Machine Learning for Molecular Design) project, which was part of the DeepMind's AI research, utilized a variety of machine learning techniques to predict the properties of molecules and design new ones. The specific methodology for the Reward Model within DeepSeek M3 is not typically detailed in a straightforward manner, as it involves complex integration of various components and approaches.

However, I can provide a general overview of the types of methods that might be used in a reward model for molecular design:

1. **Reinforcement Learning (RL):** This is a type of machine learning where an agent learns to make decisions by performing actions

Использовать RAG?

true

In DeepSeek-M3, the Reward Model (RM) utilized for reinforcement learning (RL) is based on a combination of rule-based and model-based approaches. Here's a breakdown of the methods used:

1. **Rule-Based Reward System:** For questions that can be validated using specific rules, a rule-based reward system is employed. This system is particularly useful for problems with deterministic solutions, such as math problems with known results. The model is trained to provide feedback that adheres to these predefined rules, ensuring that the responses are correct and formatted appropriately.

2. **Model-Based Reward System:** For questions with free-form ground-truth answers, the reward model is tasked with determining whether the response matches the expected ground-truth. This approach is applied to questions that do not have a definitive answer, such as those involving creative writing. The reward model is trained using DeepSeek-R1 SFT checkpoints, and preference data is constructed to include the chain-of-thought leading to the reward. This helps mitigate the risk of reward hacking in specific tasks.



Дальнейшее развитие системы

1. RAG:
 - a. Добавить метрики оценки RAG системы
 - b. Добавление BM25 для поиска релевантных чанков
 - c. Обработка PDF более мощной VL моделью
 2. Сервис:
 - a. Распараллеливание работы сервиса
 - b. Оптимизацию работы модели (vLLM)
 - c. Автоматизация сбора PDF
 - d. Деплой полновесного сервиса
-