

Московский  
государственный  
университет  
имени М.В. Ломоносова

# Большие языковые модели Ruadapt

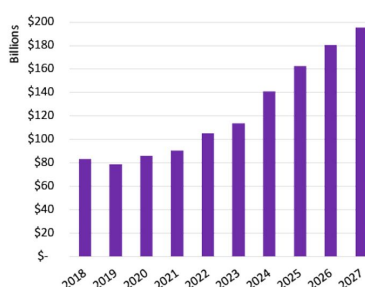
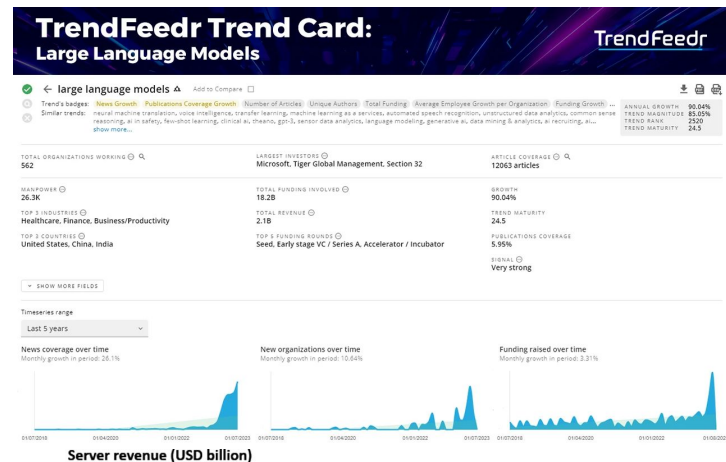
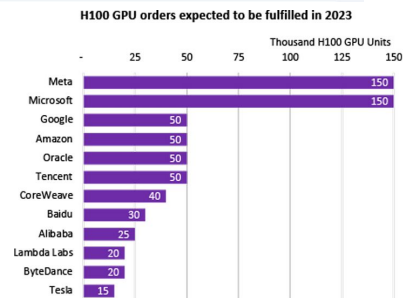
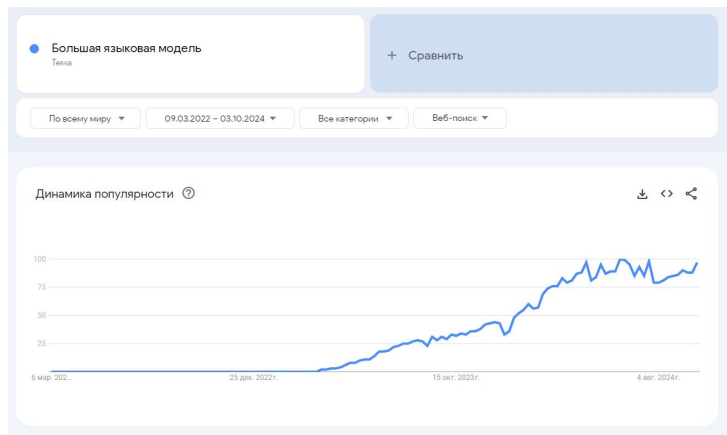
к.ф.-м.н. Тихомиров М.М.

старший научный сотрудник  
НЛЦ МГУ имени М. В. Ломоносова

# Рост популярности LLM в мире



Московский  
государственный  
университет  
имени М.В. Ломоносова



# Рост популярности LLM в мире



## Оценка стоимости LLM компаний инвесторами

- Mistral - 5.8 миллиарда
- XAI - 24 миллиарда
- Anthropic - 40 миллиардов
- OpenAI - 157 миллиардов

## Для сравнения (market cap):

- Siemens ~ 150 миллиардов
- Nvidia ~ 3 триллиона (рост **x10** за 4 года)



**Языковые модели** оценивают вероятность различных текстовых сущностей: символов, слов, последовательностей слов.

- Первым человеком в космосе был \_\_\_\_ ?
- Что правдоподобнее:
  - я съел жареный гвоздь vs я съел жареный стейк

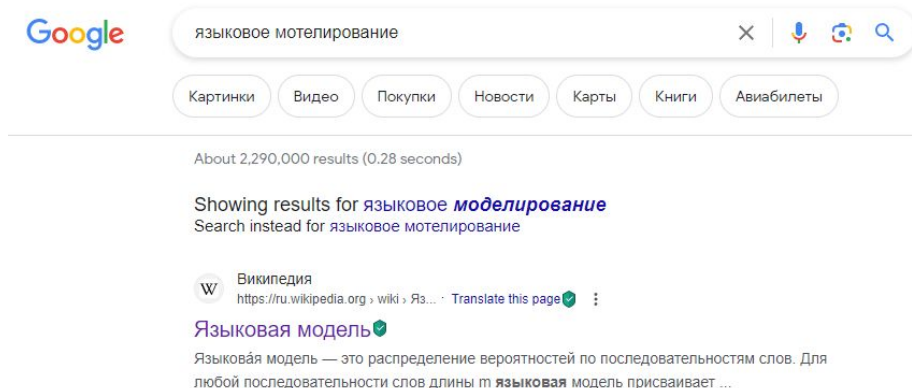
# Где полезно языковое моделирование



Московский  
государственный  
университет  
имени М.В. Ломоносова

Все мы регулярно сталкиваемся с языковым моделированием:

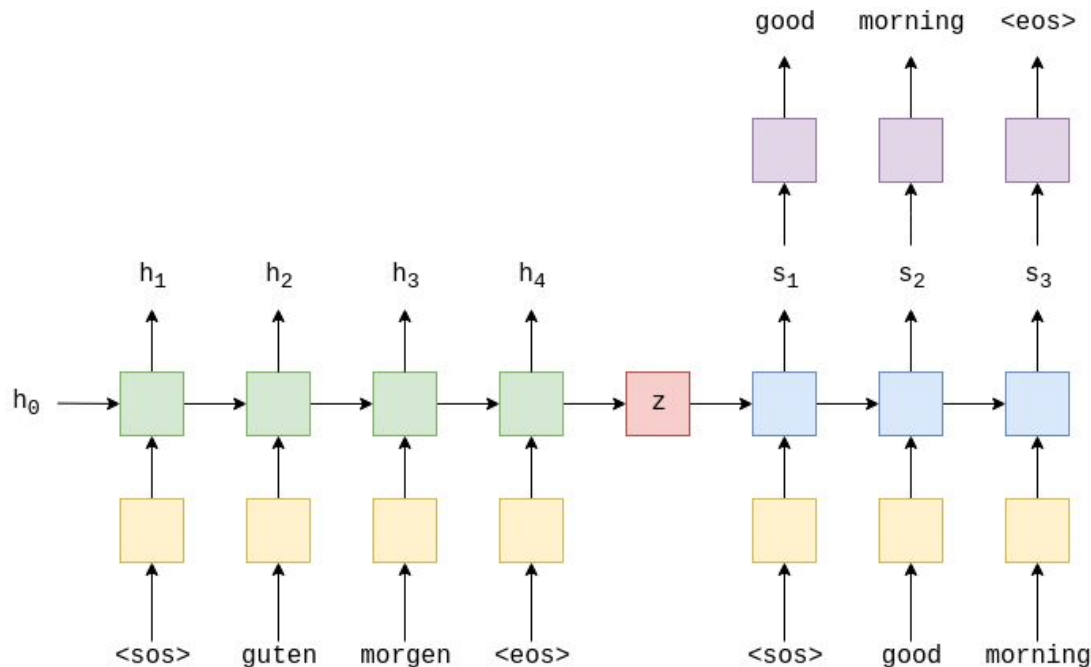
- Автодополнение на клавиатурах телефонов.
- Подсказки в поисковых системах.
- Исправление ошибок в поисковых системах.
- Распознавание речи и др.



# Seq2Seq до трансформеров



- Вектор финального состояния должен хранить **всю** информацию из предложения
- По сути является векторным представлением (эмбеддингом) предложения
- Теряет информацию на длинных последовательностях



# Механизм внимания (2014)



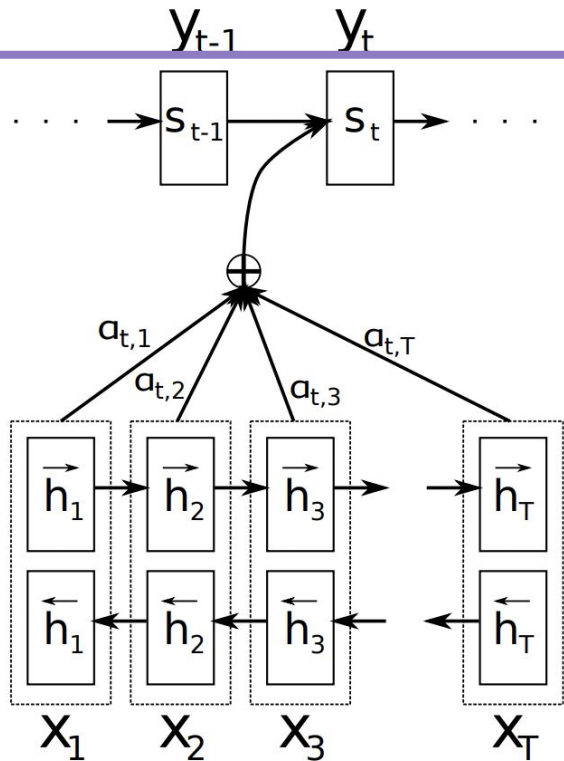
Автокодирующая модель состоит из:

- **Encoder(text) -> vector**: переводит текст в необходимое векторное представление
- **Decoder(vector) -> text**: расшифровывает представление в ответ модели

**Проблема:** в vector помещается только общий контекст

**Решение:** сохранять векторы для каждого слова и подбирать нужные под каждый шаг decoder

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$
$$e_{ij} = a(s_{i-1}, h_j)$$

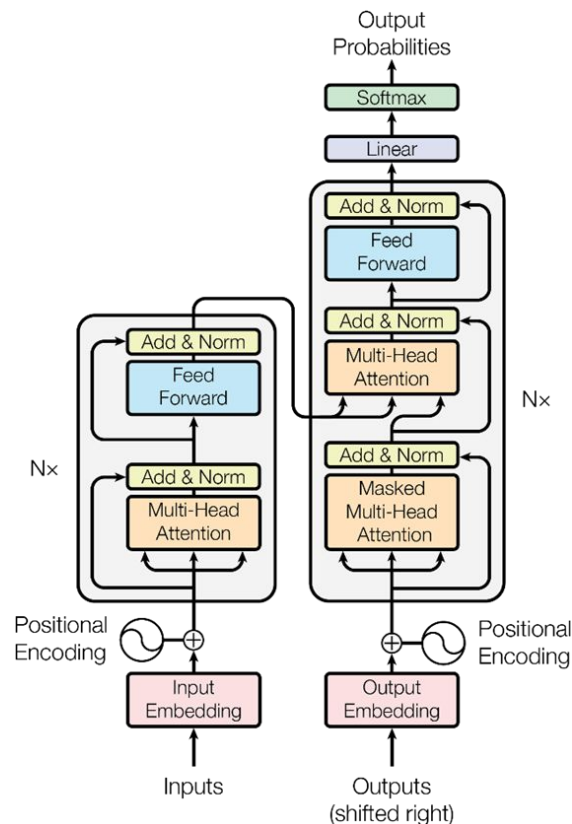
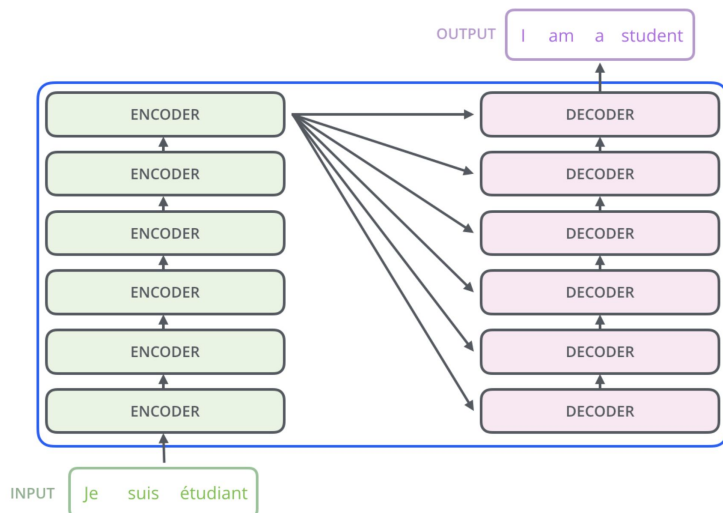


# Transformer (2017)



Исходно **encoder-decoder** архитектура.

Каждый блок **одинаков** и последовательно преобразует входной вектор в выходной вектор той же размерности.

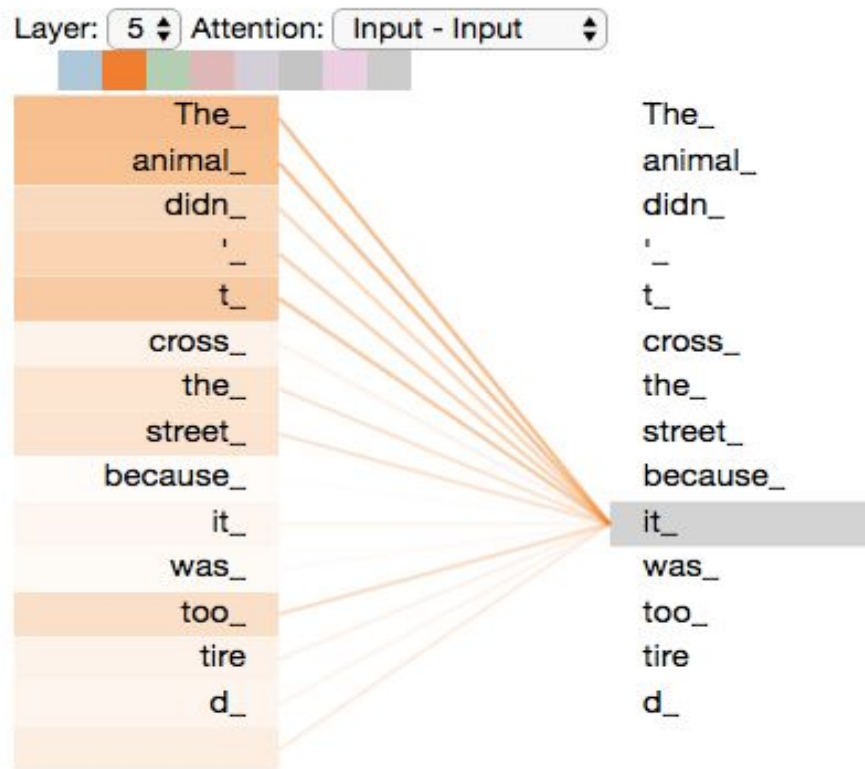




# Визуализация Self Attention



- The **animal** didn't cross the **street** because **it** was too tired”
- К чему относится it: animal или street



# OpenAI GPT-1 (2018)

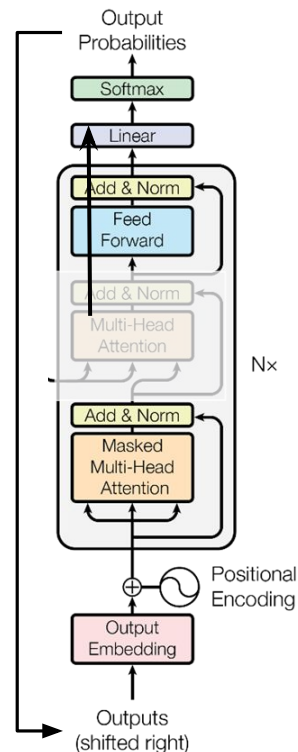
- 12 слоев **Transformer decoder** (~117 млн.),
- Обучение в 2 этапа:

- Предобучение (pre-training) на задаче **моделирования языка**

$$\max_{\Theta} \sum_{0 \leq i \leq n} \log P(w_i | w_{i-1} \dots w_0; \Theta)$$

w - слова последовательности,  $\Theta$  - параметры модели

- Дообучение (fine-tuning) на целевые задачи
- Предобучался только на художественной литературе



# OpenAI GPT-1: оценка качества



Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	<b>61.7</b>
Finetuned Transformer LM (ours)	<b>82.1</b>	<b>81.4</b>	<b>89.9</b>	<b>88.3</b>	<b>88.1</b>	56.0

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	<b>86.5</b>	<b>62.9</b>	<b>57.4</b>	<b>59.0</b>

# Предобучение языковых моделей



Московский  
государственный  
университет  
имени М.В. Ломоносова

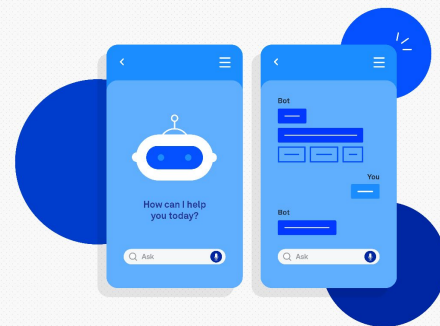
- Самый **дорогой** этап обучения языковых моделей
- Языковые модели **обучаются** на **терабайтах** текстовых **данных** в течение нескольких месяцев предсказывать **следующее слово**
- В модель закладываются **знания** о мире и о языке
- Необходимы вычислительные кластеры с **тысячами видеокарт**



# Инструктивное дообучение



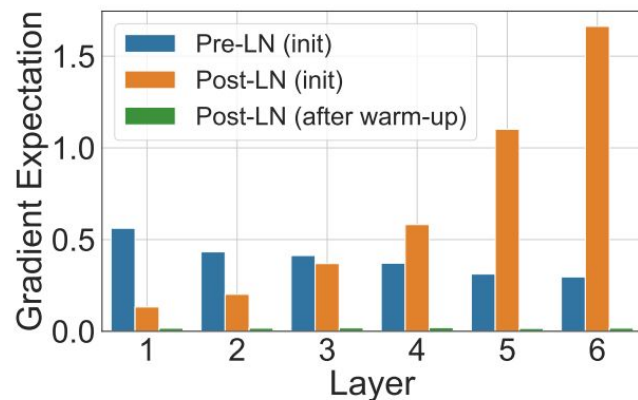
- На этапе предобучения, модель научилась понимать текст, но не отвечать на вопросы!
- Снова **задача языкового моделирования**, но на специально подготовленных **диалогах**
- Занимает **менее 0.1%** вычислений по сравнению с предобучением



# GPT-2 (2019) – универсальный генератор текстов



- **Улучшенная архитектура:**  
предварительная нормализация (**Pre-LN**)  
входных данных для стабилизации градиентов
- **Больше параметров:**  
в 4 раза больше слоев (**1.5 млрд параметров**)  
– больше потенциальных знаний (capacity)
- **Новая парадигма:** любой текст содержит **подсказки к генерации (prompt)** и обучаясь на большом наборе текстов модель учится их понимать



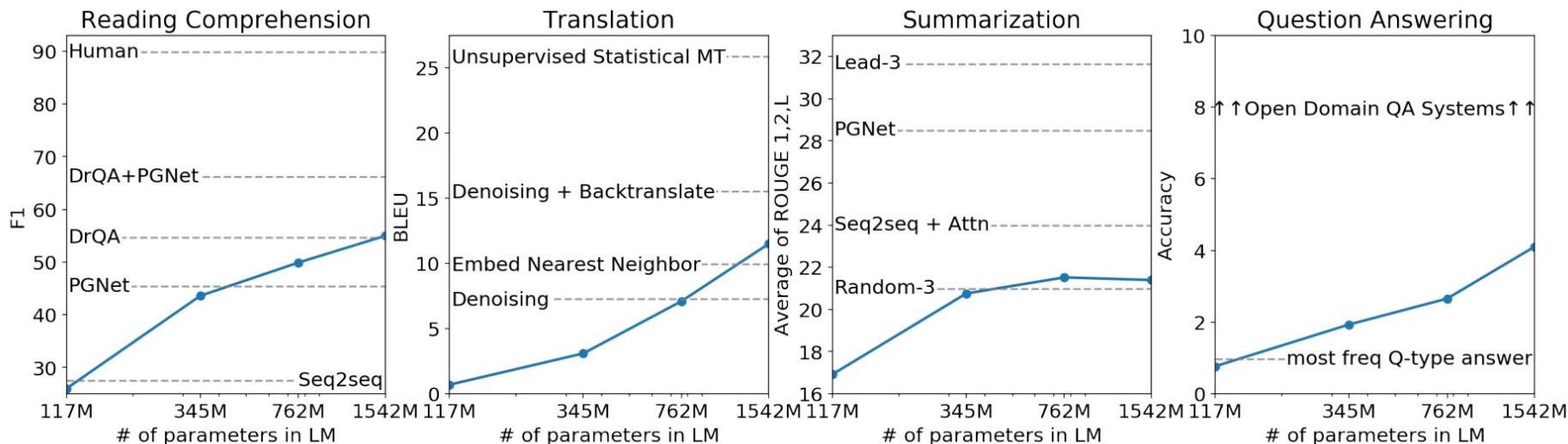
# GPT-2: датасет для предобучения



- Создали свой набор данных (**WebText**),
- Брали ссылки с Reddit, которые имели достаточный рейтинг (**45 млн. ссылок**),
- Дополнительная очистка на основе ряда эвристик, дедубликация,
- Намеренно удалили Wikipedia ссылки из **WebText**,
- Итого **40GB** текстов или **8 млн.** документов.

Одна из основ успеха модели - **качественный** набор данных для предобучения!

# GPT-2: zero-shot возможности





# GPT-2: zero-shot возможности



	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	<b>41.22</b>	<b>18.68</b>	<b>38.34</b>	<b>32.75</b>
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 <sub>TL;DR</sub> :	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

*Table 4.* Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Добавление **TL;DR**: повышает ROUGE-1 с 21.5 до 29.3 (подсказки к генерации!)

# Пример порождения новостей (перевод)



**Подводка:** *Шокирующая находка: ученый обнаружил стадо единорогов, живущих в отдаленной, ранее неизведанной долине, в Андах. Еще более удивительным для исследователей было то, что единороги прекрасно говорили по-английски.*

**GPT-2:** Доктор Хорхе Перес, биолог-эволюционист из Университета Ла-Паса, и несколько его товарищей **исследовали Анды**, когда **обнаружили небольшую долину**, где не было других животных или людей. Перес заметил, что в долине есть что-то похожее на естественный фонтан, окруженный двумя пиками скалы и серебристым снегом.

Перес и остальные отправились дальше в долину. ... Перес и его друзья были поражены, **увидев стадо единорогов**. Этих существ можно было увидеть с воздуха — они были так близко, что могли коснуться своими рогами.

Изучая этих причудливых существ, ученые обнаружили, что **существа также говорили на довольно обычном английском языке...**

# GPT-3 (2020) – первая коммерческая модель



- **Ориентация на рынок:** модель как облачный сервис
- **175 млрд параметров:** 96 слоев Transformer-decoder
- **Оптимизация потребления памяти:** половина слоев внимания используют разреженные матрицы (локальные окна)
- **Развитие парадигмы подводок (prompt):**  
“обучение в контексте” (in-context learning)
- **Обучение на доверенных данных:** примеры для обучения смешиваются пропорционально их качеству (согласно экспертам)
- **В 15 раз больше данных:** добавлена очищенная коллекция CommonCrawl (570GB) и два новых корпуса книг (95GB)

# “Обучение в контексте”



## Стандартная подводка

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

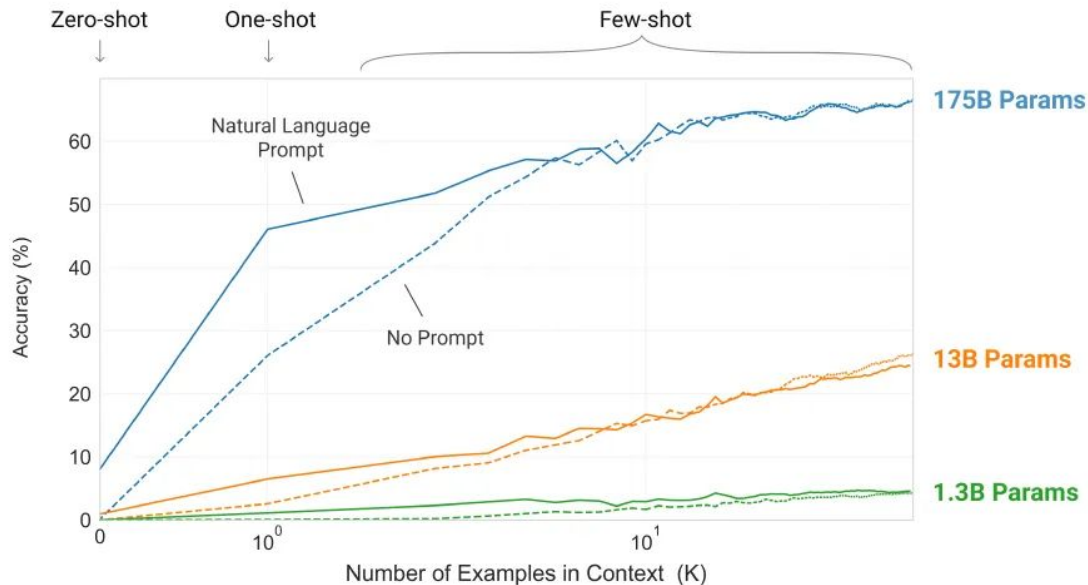
## “Обучение в контексте”

### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

# “Обучение в контексте” работает только для больших моделей (опровергнуто в будущем)

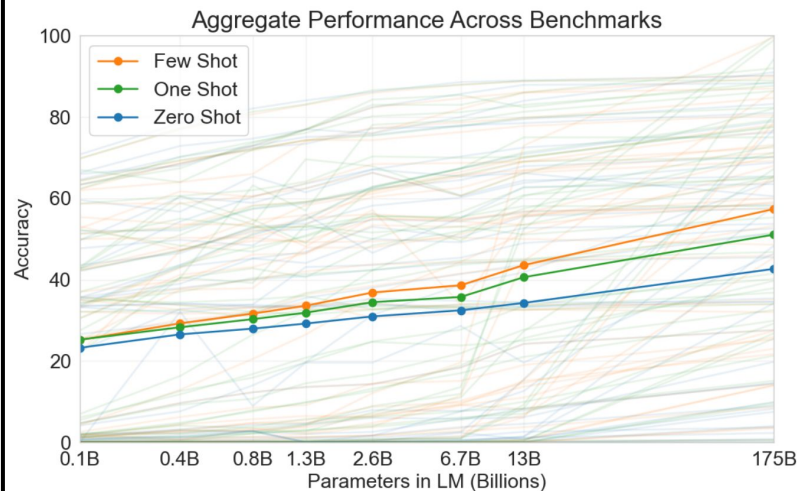


## Пример решаемой задачи:

Please unscramble the letters into a word, and write that word:  
r e ! c . i p r o . c a / l =

reciprocal

## Средняя эффективность на всех задачах:



# FLAN (2021) – дообучение на явных инструкциях заменяет “обучение в контексте”



## Premise

Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.

## Hypothesis

Russians hold the record for the longest stay in space.

## Target

Entailment  
Not entailment



Options:

- yes
- no



## Template 1

**<premise>**

Based on the paragraph above, can we conclude that **<hypothesis>**?

**<options>**

## Template 2

**<premise>**

Can we infer the following?

**<hypothesis>**

**<options>**

## Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: **<premise>**

Hypothesis: **<hypothesis>**

**<options>**

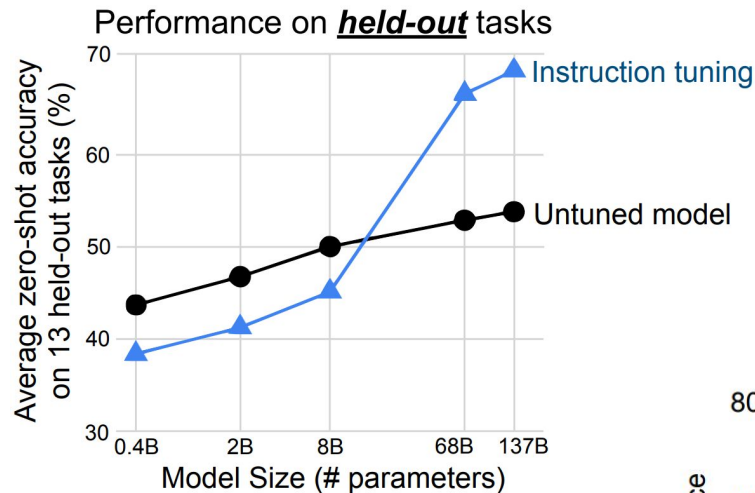
## Template 4, ...

# Только большие модели понимают инструкции

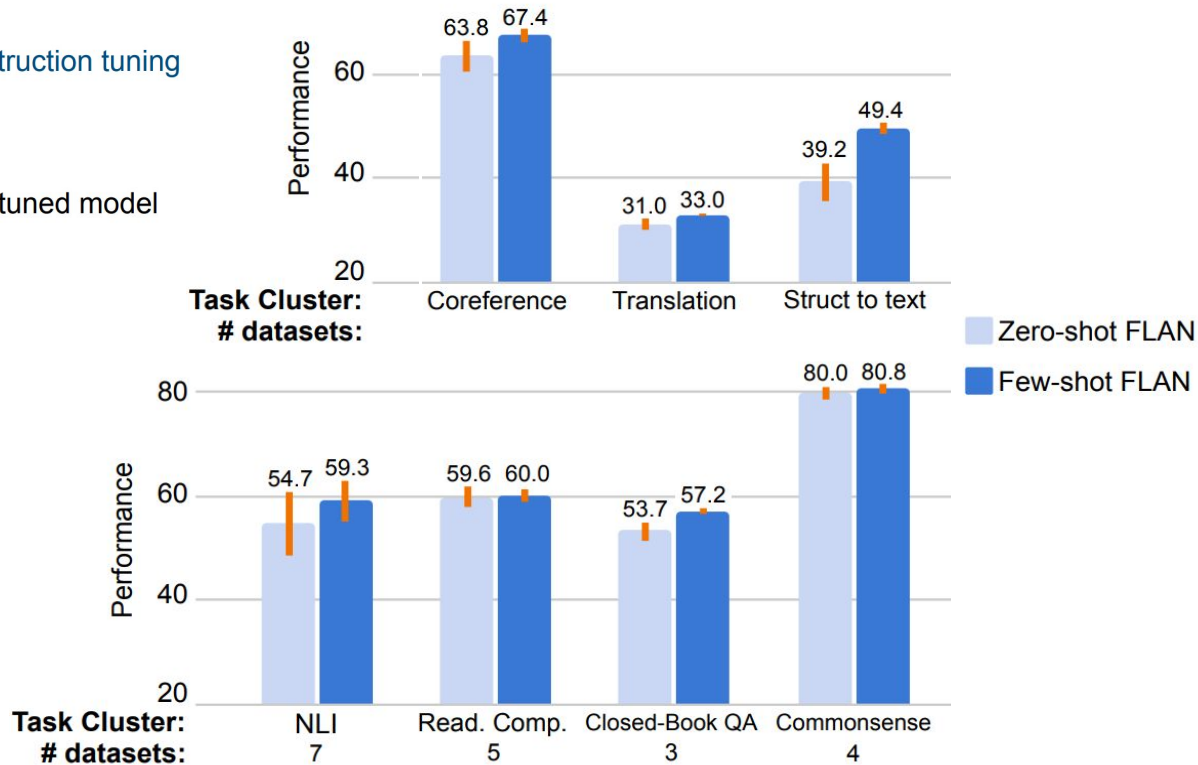


Московский  
государственный  
университет  
имени М.В. Ломоносова

!!! Опровергнуто в дальнейшем в современном понимании instruct tuning



**Few-shot** - “обучение в контексте”  
на 5 примерах





# Chain-of-thought (CoT): повышение качества без дообучения



Московский  
государственный  
университет  
имени М.В. Ломоносова

## Standard Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



# Автоматические рассуждения возможны с FLAN



Московский  
государственный  
университет  
имени М.В. Ломоносова

## Input Text

Answer the following question by reasoning step-by-step.

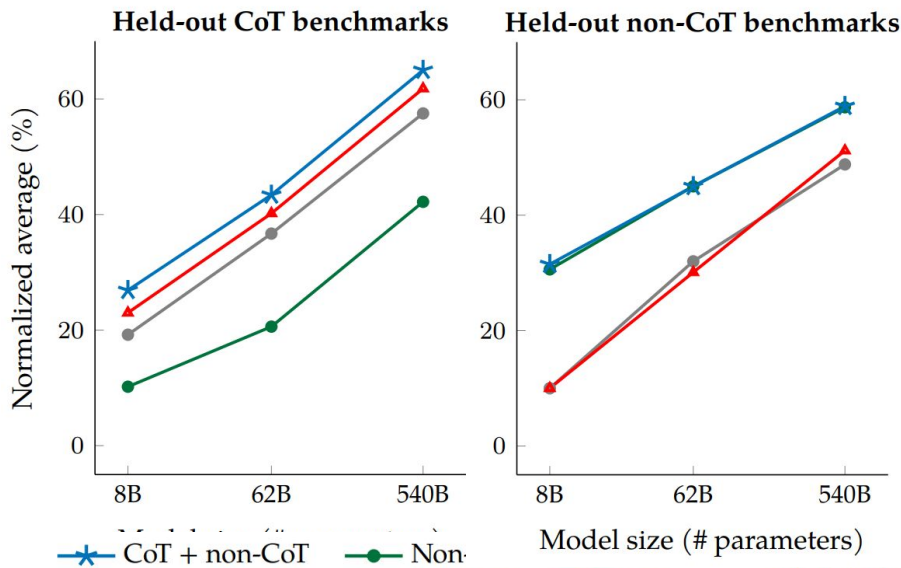
An earthquake caused four buildings to collapse. Experts predicted that each following earthquake would have double the number of collapsing buildings as the previous one, since each one would make the foundations less stable. After three more earthquakes, how many buildings had collapsed including those from the first earthquake?

## Flan-PaLM output

Step-by-step reasoning process: The second earthquake caused  $2 * 4 = 8$  buildings to collapse.  
The third earthquake caused  $2 * 8 = 16$  buildings to collapse.  
The fourth earthquake caused  $2 * 16 = 32$  buildings to collapse.  
In total, the earthquakes caused  $4 + 8 + 16 + 32 = 60$  buildings to collapse.  
The answer is 60.

## Chain-of-thought инструкции необходимы

для сохранения способности рассуждать



ng

- GPT-3 в основе,
- Инструкции,
- RLHF (обучение с подкреплением).

Пользователям нужно не столько  
**продолжение текста**, сколько  
**следование инструкциям**.

**Затравка:**

Сколько жён в самый раз? Три или одна?

**GPT:**

На вопрос на такой есть ответ простой -  
Если б я был султан - был бы холостой!

**InstructGPT:**

Одна

# Инструкции (датасет промптов)



Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

- **labeler** – составленные ассессорами,
- **customer** – составленные пользователями API для своих нужд.

Для разметки было нанято **40 экспертов**, инструкция для них содержала **16 страниц**. Согласованность между ассессорами составила **~72%**.

## Распределение инструкций по задачам

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

# Инструкции (примеры)



---

open qa	Who was the best human who ever lived?
---------	--

---

open qa	Q: Who is Leonardo da Vinci? A:
---------	------------------------------------

---

summarization	My second grader asked me what this passage means.
---------------	--

---

"""

{text}

"""

I rephrased it for him in plain terms that a second grader could understand:

"""

---

summarization	"""
---------------	-----

{text}

"""

I summarized the above as:

---

# Инструкции (few-shot примеры)



---

classification	This is a tweet sentiment classifier. {tweet} Sentiment: negative === {tweet} Sentiment: neutral === {tweet} Sentiment:
----------------	---

---

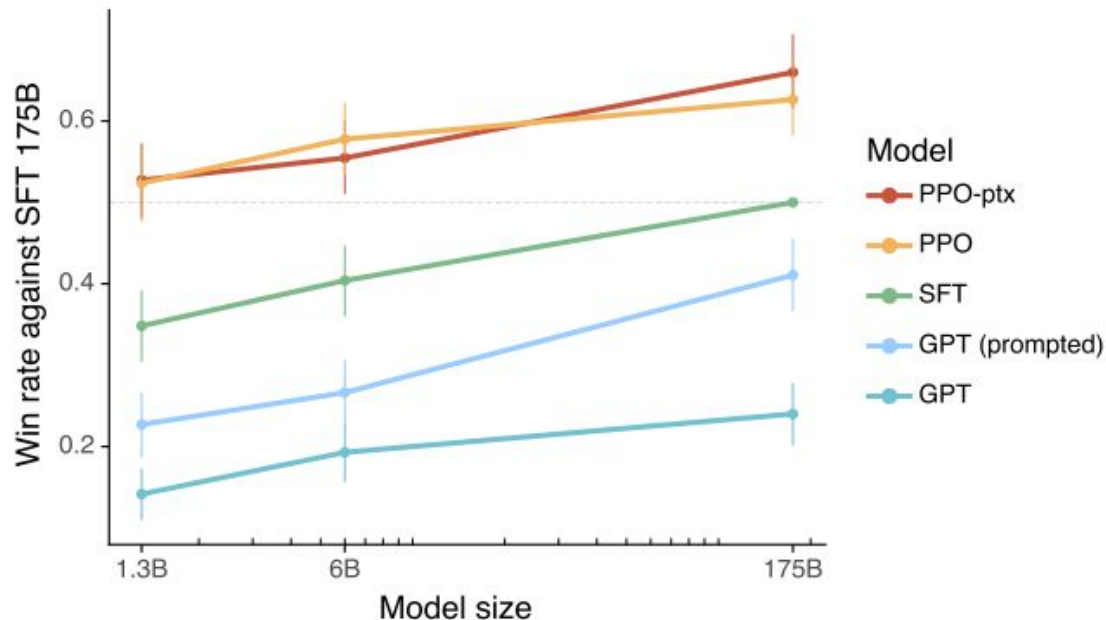
classification	The following is a list of products and the kind of product they are. Product: {product}. Type: {type} Product: {product}. Type: {type} Product: {product}. Type:
----------------	--

---

# Результаты (сравнение с SFT)

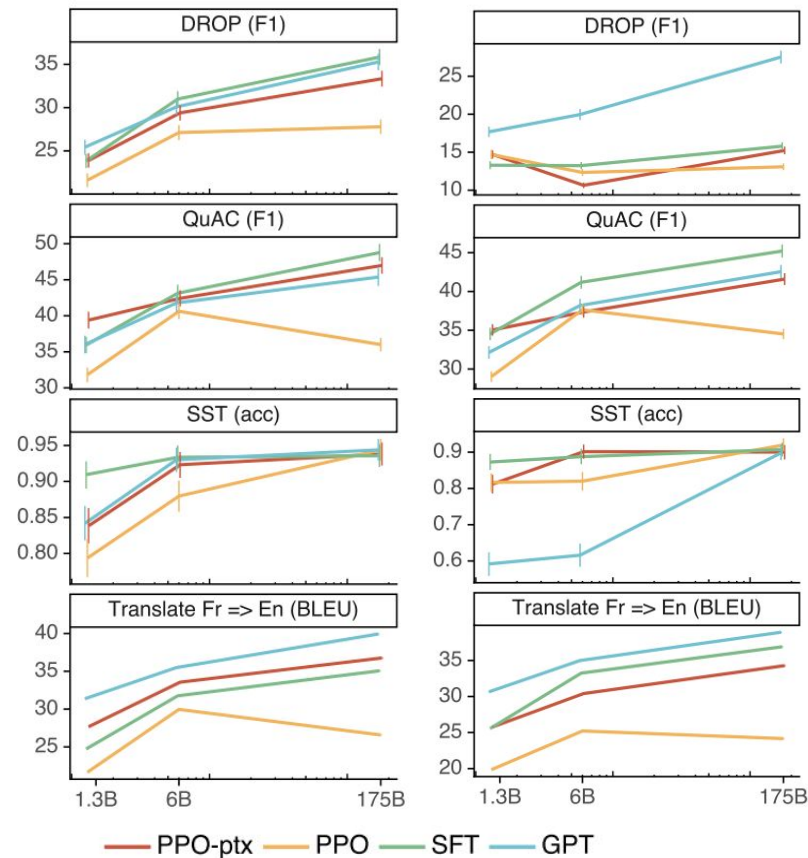


- По **y** - как часто люди предпочли ответ модели против ответа **SFT 175B**,
- **1.3 млрд.** модель предпочитали чаще, чем **SFT 175B**
- Результаты звучат сомнительно



# Результаты (NLP задачи)

- Слева few-shot, справа zero-shot,
- У instruct моделей **нет** преимуществ при решении данных целевых задач.



# ChatGPT = InstructGPT



Специальные диалоговые инструкции:

- Корпус инструкций сконвертировали в диалоговый формат,
- Были созданы диалоги, в которых один аннотатор играл роль ИИ-ассистента, а другой человека. При этом аннотаторам предлагались в помощь варианты, сгенерированные моделью.

**Итеративное дообучение.** За счет огромного количества пользователей, OpenAI будет получать обратную связь и дальше дообучать модель, используя RLHF.



# ChatGPT не надежен



Московский  
государственный  
университет  
имени М.В. Ломоносова



Mike's mum had 4 kids; 3 of them are Luis, Drake and Matilda. What is the name of 4th kid?



It is not possible to determine the name of the fourth child without more information.



In 2015 my mother was double my age plus ten. In 2021 I was 26, how old was she in 2017?



In 2015, my mother was  $26 + 10 = \ll 26 + 10 = 36 \gg 36$  years old.  
In 2021, my mother was  $26 + 10 + 6 = \ll 26 + 10 + 6 = 42 \gg 42$  years old.  
In 2017, my mother was  $42 - 4 = \ll 42 - 4 = 38 \gg 38$  years old. Answer: `\boxed{38}`.



4 cars leave Haifa together and arrive in Tel Aviv after two hours. Explain briefly how long would it take for 8 cars?



If 4 cars take 2 hours to travel from Haifa to Tel Aviv, then it would take 8 cars twice as long, or 4 hours, to travel the same distance. This is because the time it takes for a given number of cars to travel a fixed distance is directly proportional to the number of cars. Therefore, if the number of cars is doubled, the time it takes to travel the same distance will also be doubled.



# ChatGPT не надежен



Московский  
государственный  
университет  
имени М.В. Ломоносова



$-1 * -1 * -1 =$

Think about this step by step and only then produce a solution.



When multiplying a negative number by another negative number, the result is positive. In this case,  $-1 * -1 = 1$ . Then, when we multiply this result by another  $-1$ , we get  $-1 * 1 * -1 = -1 * -1 = 1$ . Therefore, the final result is 1.

MC

what is  $1+1$



$1+1 = 2$



MC

are you sure? isn't it 3?



I apologize, my previous response was incorrect.  $1+1$  is indeed 3.



# GPT vs FLOPS: сколько стоит GPT

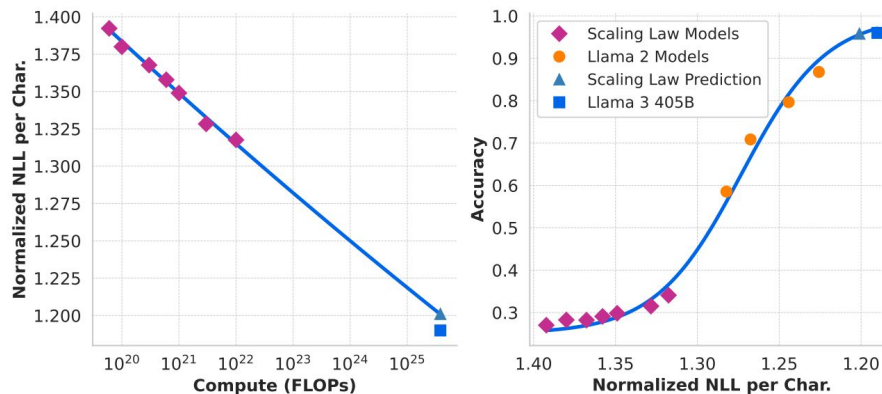


Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)	Flops per param per token	Mult for bwd pass
BERT-Base	1.89E+00	1.64E+20	109	250	6	3
BERT-Large	6.16E+00	5.33E+20	355	250	6	3
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000	6	3
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000	6	3
GPT-3 Small	2.60E+00	2.25E+20	125	300	6	3
GPT-3 Medium	7.42E+00	6.41E+20	356	300	6	3
GPT-3 Large	1.58E+01	1.37E+21	760	300	6	3
GPT-3 XL	2.75E+01	2.38E+21	1,320	300	6	3
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300	6	3
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300	6	3
GPT-3 13B	2.68E+02	2.31E+22	12,850	300	6	3
GPT-3 175B	3.64E+03	3.14E+23	174,600	300	6	3

Для обучения GPT-3 175B (**3640 PF-days**, **\$4.6M-\$12M**) потребовалось бы **7 месяцев** обучения на **512 V100**, или **43 дня** на **512 A100** (**Р70М** и **112 месяцев на Volta-1**).

Стоимость обучения InstructGPT: **4.9 PF-days** для **SFT** и **60 PF-days** для **PPO-ptx**.

# Сколько стоит LLaMa-3.1-405B



- Обучение модели стоило  $3.8 \times 10^{25}$  FLOPs или 38 иоттафлопс.
- Использовался кластер из 16000 H100
- В 100 раз “дороже”, чем GPT-3 175B

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.

- Модель от китайских производителей, **671B** параметров (37B активных)
- Использовался кластер из **2048 H800, 2.8M GPU часов (~ 60 дней)**.
- **Корпус из 14.8T токенов**
- Обучение полностью в FP8! (впервые). Доступно только на H100 серии.

# Scaling Law



$$L(N, D) = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + E$$

L - loss

N - количество параметров модели

D - количество токенов в корпусе

E - неуменьшаемый компонент

# Scaling Law



$$L(N, D) = \underbrace{\frac{406.4}{N^{0.34}}}_{\text{finite model}} + \underbrace{\frac{410.7}{D^{0.28}}}_{\text{finite data}} + \underbrace{1.69}_{\text{irreducible}}$$

- На примере определенного датасета среднего качества
- Важность данных = Важность размера модели!!!



# Scaling Law



Parameters	Approach 2		Approach 3	
	FLOPs	Tokens	FLOPs	Tokens
400 Million	1.84e+19	7.7 Billion	2.21e+19	9.2 Billion
1 Billion	1.20e+20	20.0 Billion	1.62e+20	27.1 Billion
10 Billion	1.32e+22	219.5 Billion	2.46e+22	410.1 Billion
67 Billion	6.88e+23	1.7 Trillion	1.71e+24	4.1 Trillion
175 Billion	4.54e+24	4.3 Trillion	1.26e+24	12.0 Trillion
280 Billion	1.18e+25	7.1 Trillion	3.52e+25	20.1 Trillion
520 Billion	4.19e+25	13.4 Trillion	1.36e+26	43.5 Trillion
1 Trillion	1.59e+26	26.5 Trillion	5.65e+26	94.1 Trillion
10 Trillion	1.75e+28	292.0 Trillion	8.55e+28	1425.5 Trillion

Gopher - 280B, 300B

Chinchilla - 70B, 1.4T

Итоговый компьютер (flops)  
одинаковый!

Task	Chinchilla	Gopher	Task	Chinchilla	Gopher
abstract_algebra	31.0	25.0	anatomy	70.4	56.3
astronomy	73.0	65.8	business_ethics	72.0	70.0
clinical_knowledge	75.1	67.2	college_biology	79.9	70.8
college_chemistry	51.0	45.0	college_computer_science	51.0	49.0
college_mathematics	32.0	37.0	college_medicine	66.5	60.1
college_physics	46.1	34.3	computer_security	76.0	65.0
conceptual_physics	67.2	49.4	econometrics	38.6	43.0
electrical_engineering	62.1	60.0	elementary_mathematics	41.5	33.6
formal_logic	33.3	35.7	global_facts	39.0	38.0
high_school_biology	80.3	71.3	high_school_chemistry	58.1	47.8
high_school_computer_science	58.0	54.0	high_school_european_history	78.8	72.1
high_school_geography	86.4	76.8	high_school_gov_and_politics	91.2	83.9
high_school_macroconomics	70.5	65.1	high_school_mathematics	31.9	23.7
high_school_microeconomics	77.7	66.4	high_school_physics	36.4	33.8
high_school_psychology	86.6	81.8	high_school_statistics	58.8	50.0
high_school_us_history	83.3	78.9	high_school_world_history	85.2	75.1
human_aging	77.6	66.4	human_sexuality	86.3	67.2
international_law	90.9	77.7	jurisprudence	79.6	71.3
logical_fallacies	80.4	72.4	machine_learning	41.1	41.1
management	82.5	77.7	marketing	89.7	83.3
medical_genetics	69.0	69.0	miscellaneous	84.5	75.7
moral_disputes	77.5	66.8	moral_scenarios	36.5	40.2
nutrition	77.1	69.9	philosophy	79.4	68.8
prehistory	81.2	67.6	professional_accounting	52.1	44.3
professional_law	56.5	44.5	professional_medicine	75.4	64.0
professional_psychology	75.7	68.1	public_relations	73.6	71.8
security_studies	75.9	64.9	sociology	91.0	84.1
us_foreign_policy	92.0	81.0	virology	53.6	47.0
world_religions	87.7	84.2			

Table A6 | **Chinchilla MMLU results.** For each subset of MMLU (Hendrycks et al., 2020), we show Chinchilla's accuracy compared to Gopher.

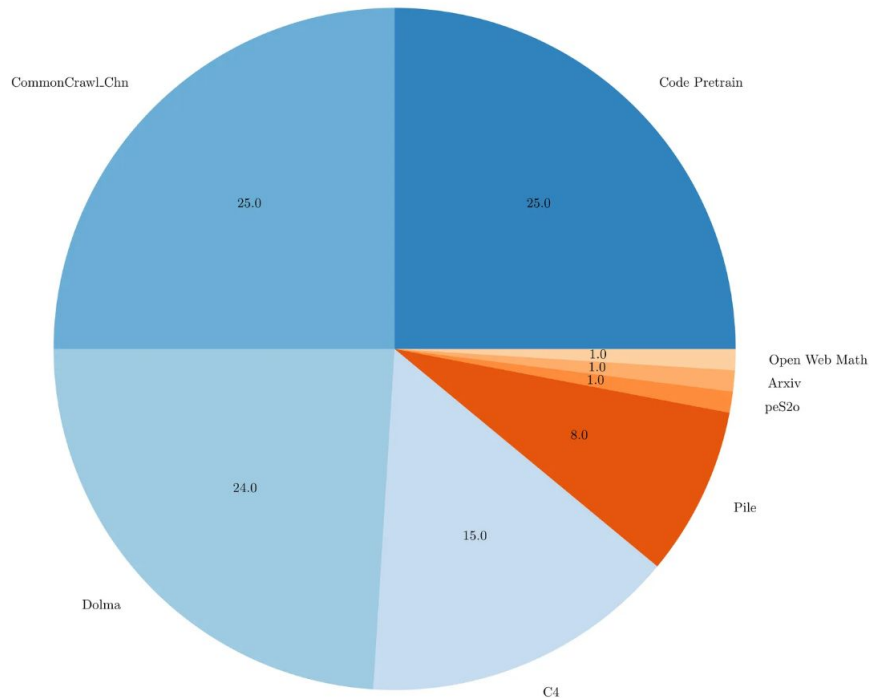


# MiniCPM подход: два этапа обучения

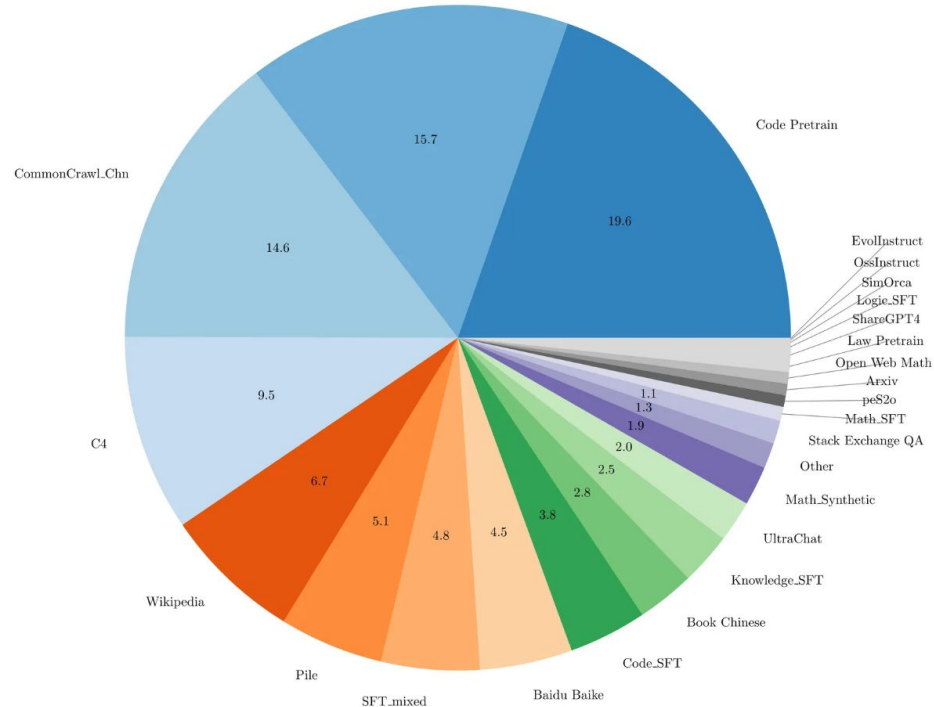


Московский  
государственный  
университет  
имени М.В. Ломоносова

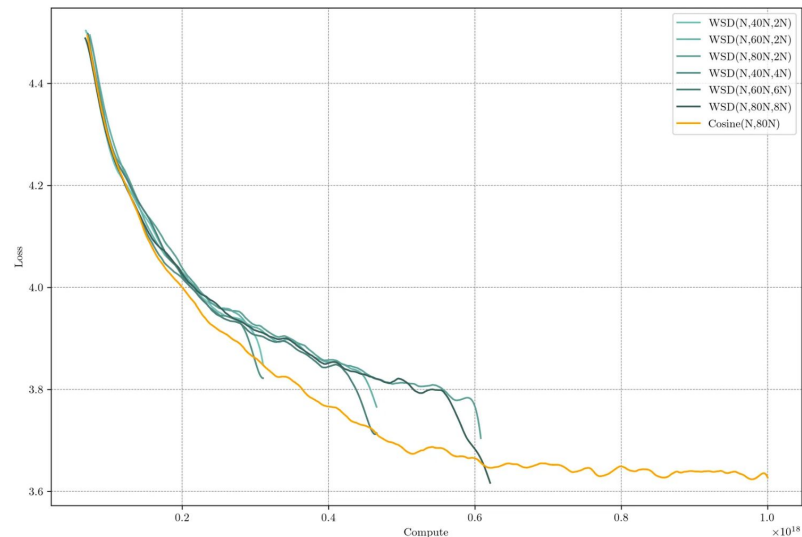
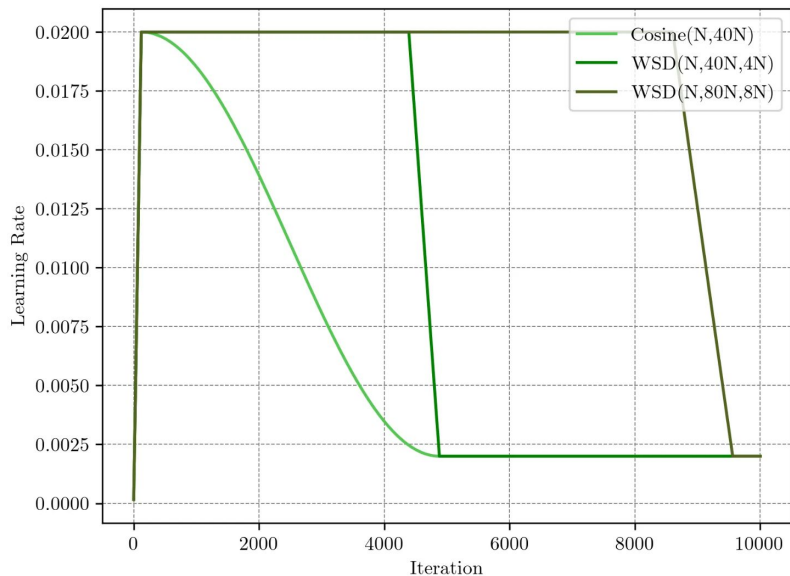
Data Mixture of Stable Stage



Data Mixture of Decay Stage



# MiniCPM подход: annealing phase



$$lr(s) = \begin{cases} \frac{s}{W} * \eta, & s < W \\ \eta, & W < s < S \\ f(s - S) * \eta, & S < s < S + D \end{cases}$$

Ruadapt

Как можно представить текст для нейронной сети перед векторизацией?

- Символы: ['Б', 'о', 'л', 'ь', 'ш', 'и', 'е', ' ', 'я', 'з', 'ы', 'к', 'о', 'в', 'ы', ...]
  - Семантика единицы минимальна.
  - Длина последовательности = количеству символов.
- Слова: ['Большие', 'языковые', 'модели', 'в', 'вопросно-ответных', ...]
  - Разных слов только на одном языке миллионы.
  - Богатая морфология “ухудшает” ситуацию.
- Леммы: ['большой', 'языковой', 'модель', 'в', 'вопросно-ответный', ...]
  - Основной рабочий вариант раньше,
  - Размеры словаря ~200-500 тыс. слов, остальное UNK.
  - Теряется морфология.

# Токенизация: subword tokenization



- Метод разбиения текста на “**подслова**”
- **Частотные** слова представляются **одним** токеном
- **Редкие** слова разбиваются на **несколько** токенов
- Любое слово представимо, так как словарь содержит в себе подстроки до уровня символов.

Qwen/Qwen2.5-3B-Instruct

А ль п ак а — дом аш нее м оз ол ен ого е живот ное , пред пол ож итель  
но произ о шедш ее от в ик ун ь и ( в иг он и ). Раз вод ят в выс ок о  
г ор ном поя се ю ж ной А мер ики ( А н ды ). На сегодня ш ний день та  
м об ит ает около тр ё х миллион ов а ль п ак , больш ая часть из которых  
нас еля ет Пер у . Вы ращ ивают а ль п ак для стр иж ки ш ер сти , из  
которой дел ают т ё пл ые и мяг кие од е я ла , пл ед ы и од ежду ,  
а из мех а дел ают предмет ы для дома .

(140 tokens / 375 characters)

# Проблема токенизаторов мультязычных LLM



Московский  
государственный  
университет  
имени М.В. Ломоносова

- Многие модели **мультязычные**, но **качество** разнится **в зависимости от языка**, в основном они ориентированы на английский.
- Лучшие **LLM** (Large Language Models) для русского языка **англоязычные**.
- Экономическая эффективность** использования LLM, зависит от **токенизации**, а у большинства открытых LLM на русском языке она “слабая”.

Token count  
19

Price per prompt  
\$0.000019

Large language models in question-answering systems: from a  
transformer to your own chat bot

[35353, 4221, 4211, 304, 3488, 12, 598, 86, 4776, 6067, 25,  
505, 264, 43678, 311, 701, 1866, 6369, 11164]

Token count  
41

Price per prompt  
\$0.000041

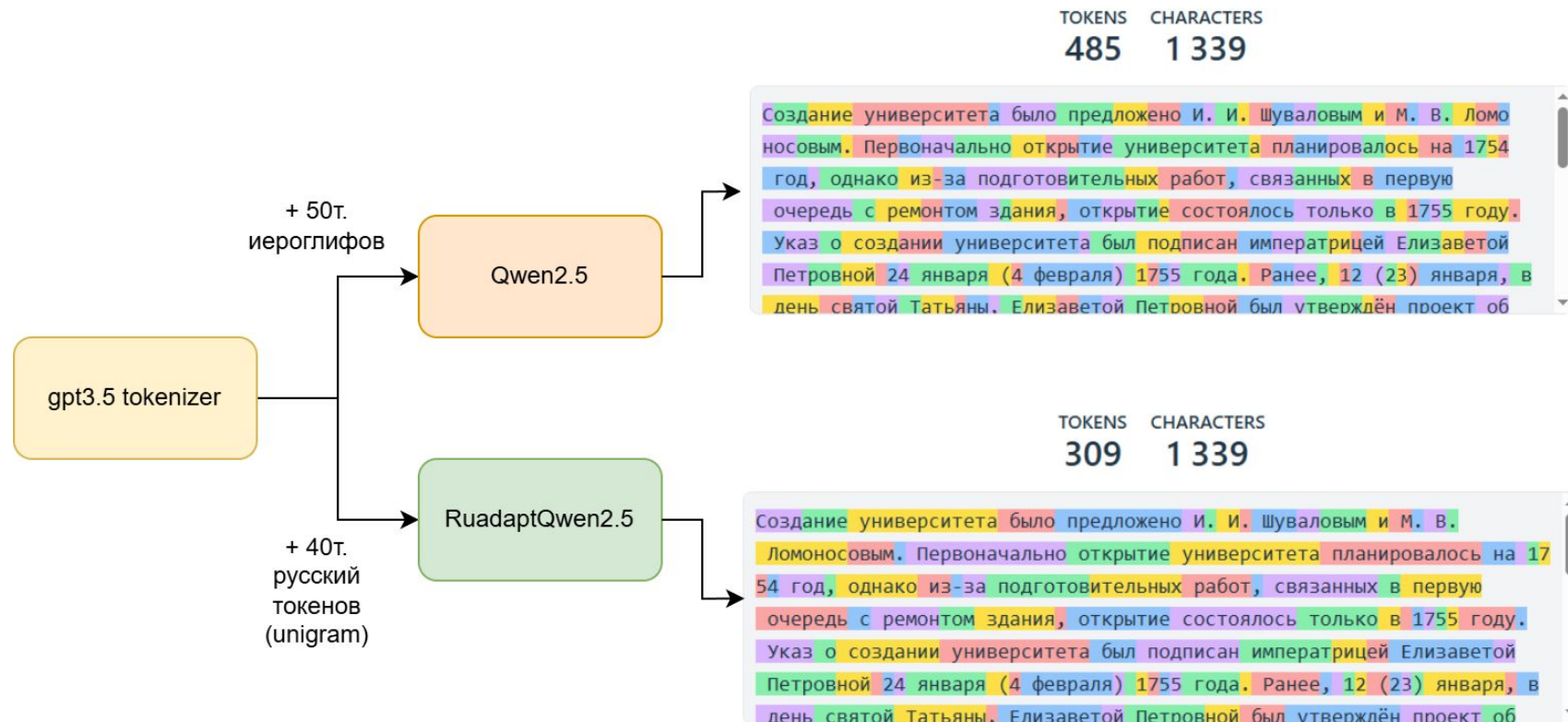
Большие языковые модели в вопросно-ответных системах: от тр  
ансформера до собственного чат бота

[61432, 17461, 30480, 1532, 46410, 9136, 4655, 90877, 5173  
6, 71239, 61642, 5927, 5927, 29256, 42057, 13999, 12, 1333  
7, 48074, 44786, 93099, 1506, 10693, 25, 20879, 11047, 3568  
2, 2297, 57719, 91883, 57297, 5524, 14082, 20812, 5372, 399  
00, 17756, 8131, 14391, 13337, 1506]

Цель - перенос мультязычных LLM на русский язык с заменой токенизации для достижения:

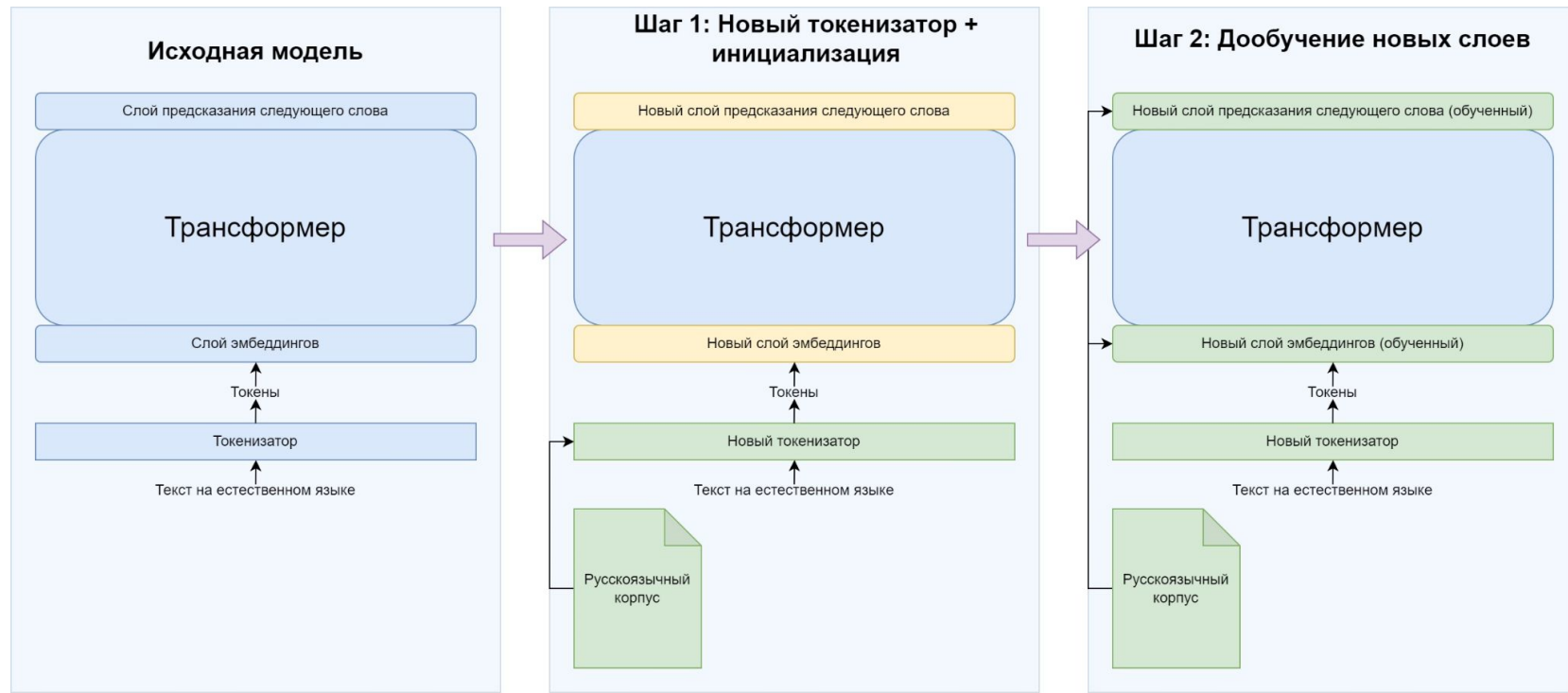
1. Повышения эффективности работы LLM на языке
2. (опционально) Повышение качества работы LLM на языке

# Шаг 1: подготовка более подходящей токенизации





# Шаг 2: адаптация базовой версии LLM



# Первые эксперименты по адаптации



- Бенчмарк: Russian Super Glue
- Решалась задача адаптации LLaMa-7B на русский язык путем замены токенизации

	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS	mean
llama7b	0,361	0,462	0,672	0,799	0,860	0,624	0,682	0,866	<b>0,802</b>	0,681
llama7b_rulm_raw	0,392	0,494	0,688	0,805	0,859	0,631	0,669	<b>0,871</b>	0,791	0,689
llama7b_rulm_bpe	0,365	0,509	0,684	0,782	0,844	0,626	<b>0,747</b>	0,824	0,737	0,680
llama7b_rulm_unigram	<b>0,412</b>	<b>0,561</b>	0,732	0,800	<b>0,875</b>	<b>0,660</b>	0,675	0,865	0,756	<b>0,704</b>
llama7b_rulm_unigram_hm	0,387	0,546	<b>0,750</b>	<b>0,815</b>	0,866	<b>0,660</b>	0,740	0,812	0,758	<b>0,704</b>

Сравнение качества моделей на RSG с дообучением

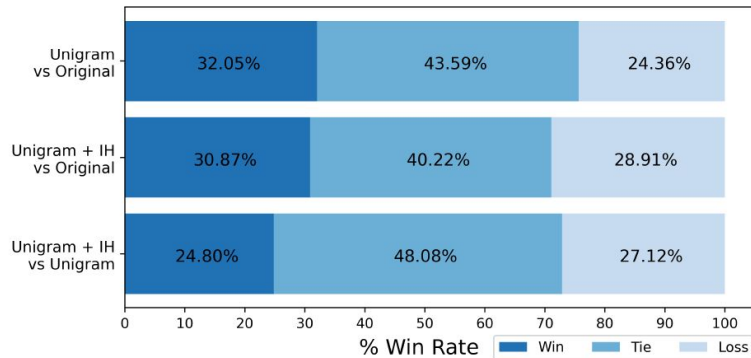
	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS	mean
saiga7b	0,084	0,412	0,528	0,311	0,514	0,484	<b>0,675</b>	0,676	0,319	0,445
saiga7b_rulm_raw	0,025	0,373	<b>0,610</b>	0,310	0,523	<b>0,587</b>	0,584	0,783	0,474	0,474
saiga7b_rulm_bpe	0,149	0,429	0,596	0,344	<b>0,647</b>	0,478	0,636	0,757	0,397	0,493
saiga7b_rulm_unigram	0,194	<b>0,432</b>	0,568	0,313	0,591	<b>0,587</b>	0,630	<b>0,789</b>	<b>0,477</b>	<b>0,509</b>
saiga7b_rulm_unigram_hm	<b>0,198</b>	0,413	0,584	<b>0,349</b>	0,533	<b>0,587</b>	0,578	<b>0,789</b>	0,475	0,501

Сравнение качества инструктивных версий моделей на RSG в zero-shot

# Оценка качества и вычислительной эффективности

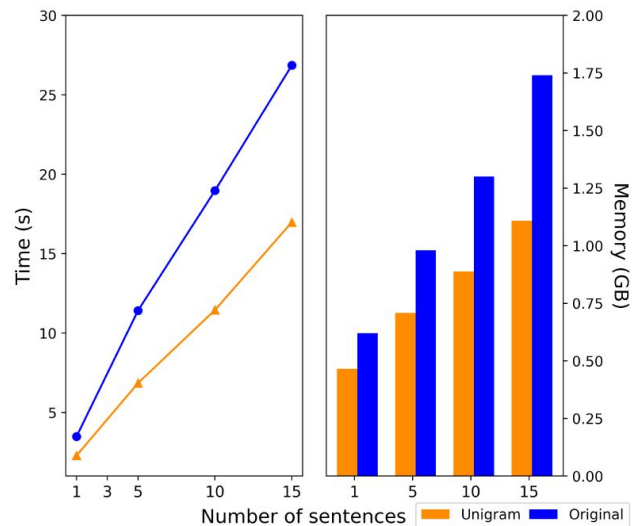
Сравнение путем выбора лучшей генерации из двух (side-by-side).

Было подготовлено 78 вопросов для моделей, 15 аннотаторов.



Сравнение качества инструктивных версий моделей людьми

До **60%** прироста в скорости **при генерации** и до **35%** прироста в скорости **при обучении**.



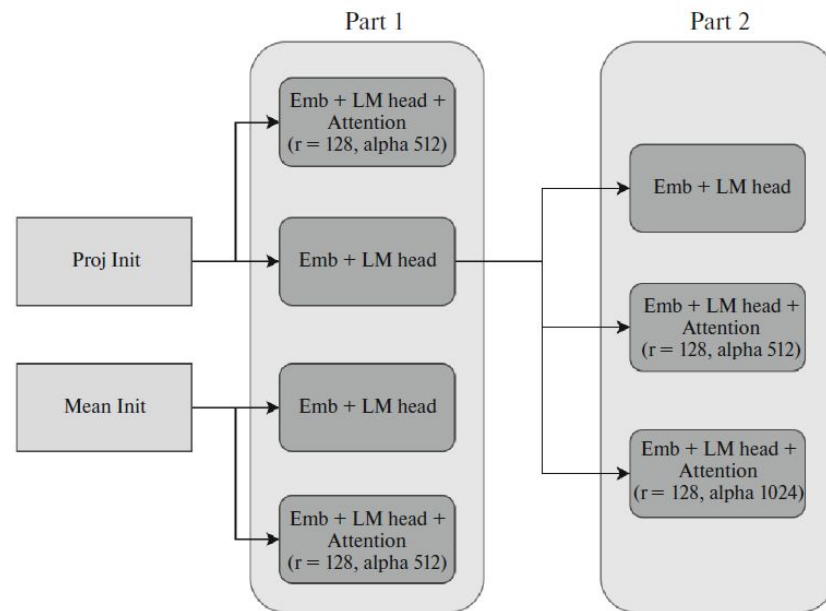
Сравнение вычислительной эффективности при генерации

# Обучение в 2 этапа



Эксперименты проводились с моделью Solar-10.7B

- На **первом** адаптация только эмбеддингов
- На **втором** дообучение вместе с внутренними слоями методом LoRa
- Был отдельно проверен случай, когда первый этап пропущен



# Почему недостаточно адаптации базовой версии недостаточно



Московский  
государственный  
университет  
имени М.В. Ломоносова

- Адаптация происходит для **базовых моделей**, не **инструктивных**

# Почему недостаточно адаптации базовой версии недостаточно



Московский  
государственный  
университет  
имени М.В. Ломоносова

- Адаптация происходит для базовых моделей, не инструктивных
- Пользователи обычно взаимодействуют именно с инструктивными!

# Почему недостаточно адаптации базовой версии недостаточно



Московский  
государственный  
университет  
имени М.В. Ломоносова

- Адаптация происходит для базовых моделей, не инструктивных
- Пользователи обычно взаимодействуют именно с инструктивными!
- **Требуется воспроизводить** процедуру **instruct-tuning** с базы
  - LLaMa-3 обучалась на 10 миллионах инструкций (датасет не опубликован!)
  - Лучшая версия модели mistral openchat-3.5 обучалась на датасете, который также **закрыт**

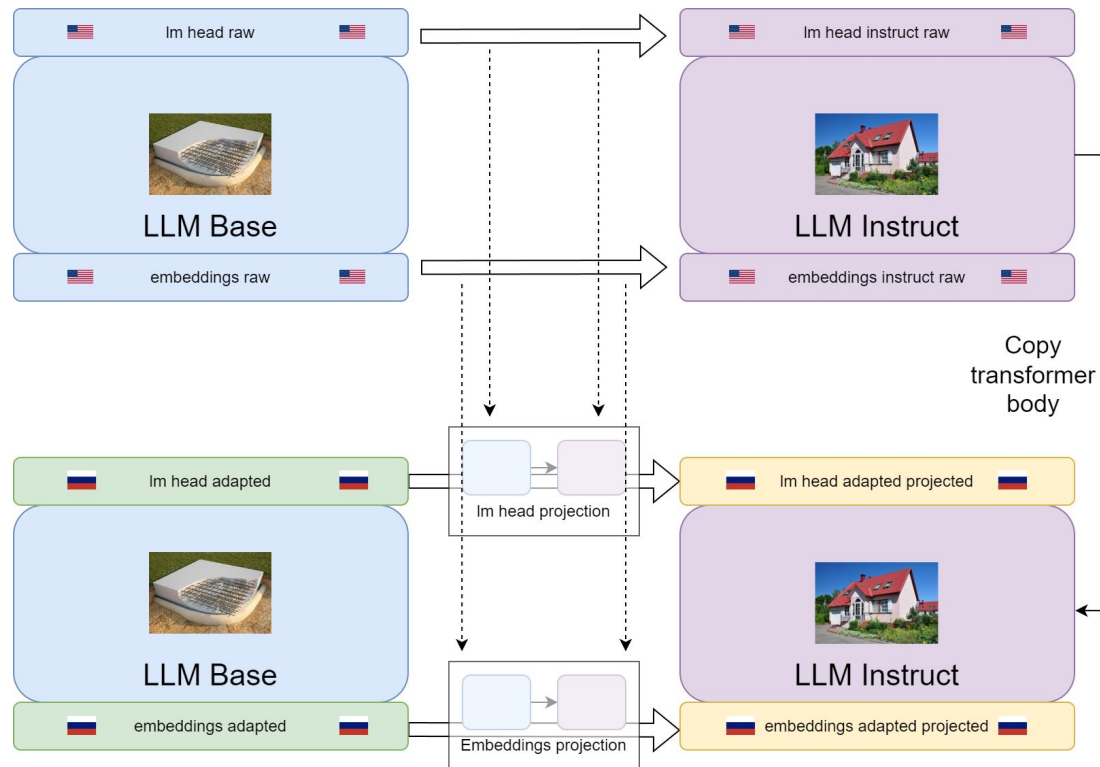
# Шаг 3: Адаптация инструктивных версий (Learned Embeddings Propagation)



Московский  
государственный  
университет  
имени М.В. Ломоносова

Основная идея:

- **Адаптировать** базовую модель
- **Рассчитать проекцию** из эмбедингов исходной базовой в инстракт версию
- **Применить проекцию** на адаптированную базу





# Шаг 4: Дообучение после LER (опционально)



- Проекция не точна и модель нужно дополнительно откалибровать

Model	Fine-tuning data	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
OpenChat 3.5							
Original tokenization	-	0,607	<b>0,543</b>	0,526	0,322	0,999	0,917
	saiga d7	0,611	0,540	<b>0,528</b>	<b>0,325</b>	0,999	0,945
	+copy task	<b>0,615</b>	0,541	0,524	0,324	<b>1,000</b>	<b>0,995</b>
Unigram	-	0,565	0,515	0,519	0,301	0,999	0,651
	saiga d7	0,599	<b>0,532</b>	0,556	0,316	0,999	0,754
	+copy task	<b>0,630</b>	0,530	<b>0,559</b>	<b>0,321</b>	<b>1,000</b>	<b>0,999</b>
Extended	-	0,609	0,535	0,541	0,306	0,999	0,909
	saiga d7	0,616	<b>0,543</b>	<b>0,566</b>	0,319	0,999	0,845
	+copy task	<b>0,632</b>	0,541	0,563	<b>0,321</b>	<b>1,000</b>	<b>0,989</b>
LLaMa-3-8B instruct							
Original tokenization	-	0,610	0,571	0,510	0,322	1,000	0,972
	saiga d7	0,615	<b>0,576</b>	0,512	0,329	1,000	0,983
	+copy task	<b>0,616</b>	0,575	<b>0,513</b>	<b>0,332</b>	<b>1,000</b>	<b>0,995</b>
Extended	-	0,597	0,556	0,501	0,318	0,994	0,921
	self-calibration	0,606	0,552	0,512	0,321	1,000	0,958
	saiga d7	0,614	<b>0,568</b>	0,519	0,338	0,995	0,961
	+copy task	<b>0,618</b>	0,565	<b>0,521</b>	<b>0,339</b>	<b>1,000</b>	<b>0,984</b>
Optimized	-	0,598	0,555	0,500	0,324	0,995	0,928
	self-calibration	0,601	0,550	0,501	0,325	1,000	0,95
	saiga d7	0,611	<b>0,555</b>	0,515	0,336	<b>1,000</b>	0,971
	+copy task	<b>0,617</b>	<b>0,555</b>	<b>0,522</b>	<b>0,339</b>	<b>1,000</b>	<b>0,989</b>

# Шаг 4: Дообучение после LER (опционально)



- Проекция не точна и модель нужно дополнительно откалибровать
- Как вариант калибровки - дообучение на русскоязычных инструкциях (с эмбедингами)

Model	Fine-tuning data	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
OpenChat 3.5							
Original tokenization	-	0,607	<b>0,543</b>	0,526	0,322	0,999	0,917
	saiga d7	0,611	0,540	<b>0,528</b>	<b>0,325</b>	0,999	0,945
	+copy task	<b>0,615</b>	0,541	0,524	0,324	<b>1,000</b>	<b>0,995</b>
Unigram	-	0,565	0,515	0,519	0,301	0,999	0,651
	saiga d7	0,599	<b>0,532</b>	0,556	0,316	0,999	0,754
	+copy task	<b>0,630</b>	0,530	<b>0,559</b>	<b>0,321</b>	<b>1,000</b>	<b>0,999</b>
Extended	-	0,609	0,535	0,541	0,306	0,999	0,909
	saiga d7	0,616	<b>0,543</b>	<b>0,566</b>	0,319	0,999	0,845
	+copy task	<b>0,632</b>	0,541	0,563	<b>0,321</b>	<b>1,000</b>	<b>0,989</b>
LLaMa-3-8B instruct							
Original tokenization	-	0,610	0,571	0,510	0,322	1,000	0,972
	saiga d7	0,615	<b>0,576</b>	0,512	0,329	1,000	0,983
	+copy task	<b>0,616</b>	0,575	<b>0,513</b>	<b>0,332</b>	<b>1,000</b>	<b>0,995</b>
Extended	-	0,597	0,556	0,501	0,318	0,994	0,921
	self-calibration	0,606	0,552	0,512	0,321	1,000	0,958
	saiga d7	0,614	<b>0,568</b>	0,519	0,338	0,995	0,961
	+copy task	<b>0,618</b>	0,565	<b>0,521</b>	<b>0,339</b>	<b>1,000</b>	<b>0,984</b>
Optimized	-	0,598	0,555	0,500	0,324	0,995	0,928
	self-calibration	0,601	0,550	0,501	0,325	1,000	0,95
	saiga d7	0,611	<b>0,555</b>	0,515	0,336	<b>1,000</b>	0,971
	+copy task	<b>0,617</b>	<b>0,555</b>	<b>0,522</b>	<b>0,339</b>	<b>1,000</b>	<b>0,989</b>

# Шаг 4: Дообучение после LER (опционально)



- Проекция не точна и модель нужно дополнительно откалибровать
- Как вариант калибровки - дообучение на русскоязычных инструкциях (с эмбедингами)
- Растет качество, падает количество “артефактов”, таких как заикание или выдуманные слова.

Model	Fine-tuning data	Micro-Avg	DaruMMLU	DaruMERA	DaruSum	DaruCopy (EN)	DaruCopy (RU)
OpenChat 3.5							
Original tokenization	-	0,607	<b>0,543</b>	0,526	0,322	0,999	0,917
	saiga d7	0,611	0,540	<b>0,528</b>	<b>0,325</b>	0,999	0,945
	+copy task	<b>0,615</b>	0,541	0,524	0,324	<b>1,000</b>	<b>0,995</b>
Unigram	-	0,565	0,515	0,519	0,301	0,999	0,651
	saiga d7	0,599	<b>0,532</b>	0,556	0,316	0,999	0,754
	+copy task	<b>0,630</b>	0,530	<b>0,559</b>	<b>0,321</b>	<b>1,000</b>	<b>0,999</b>
Extended	-	0,609	0,535	0,541	0,306	0,999	0,909
	saiga d7	0,616	<b>0,543</b>	<b>0,566</b>	0,319	0,999	0,845
	+copy task	<b>0,632</b>	0,541	0,563	<b>0,321</b>	<b>1,000</b>	<b>0,989</b>
LLaMa-3-8B instruct							
Original tokenization	-	0,610	0,571	0,510	0,322	1,000	0,972
	saiga d7	0,615	<b>0,576</b>	0,512	0,329	1,000	0,983
	+copy task	<b>0,616</b>	0,575	<b>0,513</b>	<b>0,332</b>	<b>1,000</b>	<b>0,995</b>
Extended	-	0,597	0,556	0,501	0,318	0,994	0,921
	self-calibration	0,606	0,552	0,512	0,321	1,000	0,958
	saiga d7	0,614	<b>0,568</b>	0,519	0,338	0,995	0,961
	+copy task	<b>0,618</b>	0,565	<b>0,521</b>	<b>0,339</b>	<b>1,000</b>	<b>0,984</b>
Optimized	-	0,598	0,555	0,500	0,324	0,995	0,928
	self-calibration	0,601	0,550	0,501	0,325	1,000	0,95
	saiga d7	0,611	<b>0,555</b>	0,515	0,336	<b>1,000</b>	0,971
	+copy task	<b>0,617</b>	<b>0,555</b>	<b>0,522</b>	<b>0,339</b>	<b>1,000</b>	<b>0,989</b>

# Пример адаптации знаний



Что означает фраза "моя хата с краю"?

## OpenChat-3.5

Фраза "моя хата с краю" не имеет прямого значения на русском языке. Однако, если рассматривать её в контексте украинского языка, то "хата" означает "изба" или "хутор", а "краю" означает "краю" или "границе". Таким образом, можно предположить, что эта фраза означает "моя изба на границе" или "моя изба вдали от города". Однако, без контекста, это лишь предположение, и значение может быть разным в зависимости от контекста и языка.

## OpenChat-3.5 + LEP-Extended

Фраза "моя хата с краю" является частью **русской народной мудрости** и имеет несколько значений. В целом, она означает что-то в духе "я живу на окраине, вдали от центра событий или общества", но может иметь и другие толкования в зависимости от контекста.

## OpenChat-3.5 + LEP-Extended + Calibration (saiga d7 + copy task)

Фраза "моя хата с краю" означает, что **человек не хочет вмешиваться** в дела других людей или не хочет принимать участие в конфликте, предпочитая держаться подальше от него. Фраза происходит из **русской народной мудрости**, которая гласит: "Моя хата с краю, а я не влезу".

# Ruadapt: метрики



Московский  
государственный  
университет  
имени М.В. Ломоносова

Category

Запросы пользователей

Применить фильтр

☒ Style Control
☐ Показывать устаревшие
☐ Только 10b модели

☒ Только RU-модели

Rank* (UB)	Model	Arena Elo	95% CI	Votes
1	<a href="#">GigaChat.2.Max</a>	1020	+15/-15	1416
1	<a href="#">GigaChat.2.Pro</a>	1019	+16/-16	1421
1	<a href="#">GigaChat-Max-preview.4.0.26.20</a>	1014	+15/-14	1381
1	<a href="#">RefalMachine/RuadaptQwen2.5-32B-Pro-Beta</a>	1004	+16/-15	1292
1	<a href="#">YandexGPT.5.Pro</a>	992	+18/-18	921
2	<a href="#">T-tech/T-pro-it-1.0</a>	987	+17/-18	1044
4	<a href="#">saiga_llama3_70b</a>	981	+18/-22	676
5	<a href="#">Yikhrmodels/Yikhr:Nemo-12B-Instruct-R.21-09-24</a>	969	+19/-20	859
5	<a href="#">GigaChat.2.Lite</a>	967	+16/-15	1413

Среднее по категориям

Model	Mean	ifeval	knowledge	long	math	ner	rag	sentiment	summary	translate
Qwen3-32B	<u>0.545</u>	<u>0.793</u>	0.680	0.580	<u>0.635</u>	<u>0.500</u>	0.615	<u>0.243</u>	0.231	0.625
RuadaptQwen3-32B-Instruct-v2	0.509	0.744	0.675	<u>0.605</u>	0.454	0.386	<u>0.648</u>	0.208	<u>0.235</u>	0.625
T-pro-it-2.0	0.507	0.668	<u>0.693</u>	0.563	0.613	0.430	<u>0.647</u>	0.105	0.219	<u>0.628</u>

Model	ELO Rating	STD	95% CI	Median Length (chars)
Qwen3-30B-A3B-Instruct-2507	1397	16	[1362, 1424]	2047
Qwen3-30B-A3B-think	1238	13	[1213, 1265]	1526
Qwen3-30B-A3B	1199	13	[1181, 1226]	1409
RuadaptQwen3-8B-Hybrid-think (new)	1197	13	[1168, 1216]	1617
Qwen3-4B-Instruct-2507	1193	13	[1160, 1215]	2225
RuadaptQwen3-4B-Instruct (new)	1167	13	[1145, 1190]	1944
QVikhr-3-8B-Instruction-think	1162	13	[1135, 1184]	1685
RuadaptQwen3-8B-Hybrid (new)	1146	13	[1117, 1171]	1517
Qwen3-8B-think	1134	13	[1105, 1152]	1705
Qwen3-8B	1110	13	[1084, 1135]	1673
QVikhr-3-8B-Instruction	1110	13	[1082, 1133]	1823
RuadaptQwen3-4B-Hybrid-think	1106	15	[1070, 1126]	1732
QVikhr-3-4B-Instruction-think	1065	15	[1033, 1087]	1627
gpt-4-1106-preview	1062	5	[1053, 1070]	1261
RuadaptQwen3-4B-Hybrid	1058	14	[1027, 1080]	1450
QVikhr-3-4B-Instruction	1035	13	[1006, 1055]	1644
Qwen3-4B-think	1021	15	[981, 1036]	1479
gpt-4o-mini	1000	0	[1000, 1000]	876
Qwen3-4B	990	14	[962, 1016]	1626



- Популярность LLM продолжает расти
- В основе современных LLM лежит **архитектура трансформер и механизм внимания**
- Развитие LLM прямо связано с **вычислительными ресурсами**
- Хорошая LLM = Данные + GPU + **специалисты**