

Article review

Pre-Training Small Base LMs With Fewer Tokens

Presented by **Nadezhda Anisimova**

May 2024



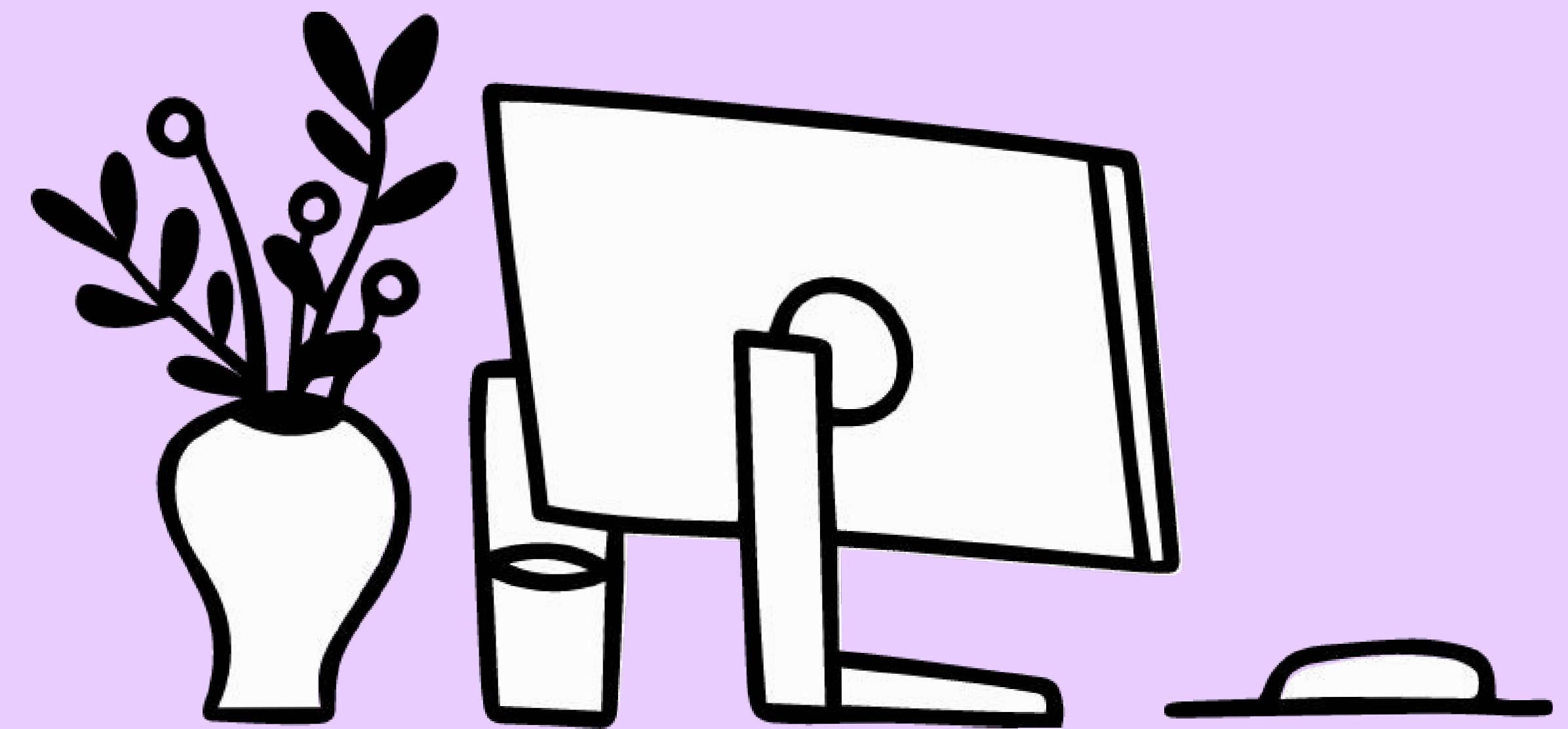


Или рецепт, как приготовить свою маленькую LM, не сильно потеряв в качестве

Статья

Репозиторий

Предисловие



Какие Этапы Превращают Простую Нейронку В LLM?

Pre-Training

Процесс обучения модели на огромном корпусе текста (миллиарды слов), чтобы модель выучила структуру языка, грамматику, общие факты о жизни.

Сравниво с обучением ребенка другому языку, посредством чтения большого количества литературы, без глубокого понимания определенных тем

Fine-Tuning

Процесс идет после pre-training, направлен на дообучение модели на узкую специализацию для определенных задач.

Сравниво с обучением ребенка определенной области, например биологии, уже после того, как им был изучен сам язык

Данные

Огромный корпус (например Википедия)

Меньший датасет под определенную задачу (например, юридические тексты)

Pre-Training

Процесс обучения языковых моделей на огромном корпусе текста (миллиарды страниц), чтобы модель выучила структуру языка, грамматику, лексику и факты о жизни.

Сравниво с обучением ребенка другому языку, посредством чтения большого количества литературы, без специального изучения определенных тем

Данные

Огромный корпус текста (например Википедия)

Предобучение требует

Мощные вычислительные ресурсы

Огромный и качественный датасет

Обычно для LLM **доступны только веса**, но не полный датасет для pre-train'a

В рамках статьи рассмотрим только pre-training

Зачем Нужна Small LM?

Маленькие языковые модели обеспечивают **высокую адаптивность и быстроту реагирования**, что важно для приложений в реальном времени. Их меньший размер **снижает задержку**, делая их идеальными для обслуживания клиентов через искусственный интеллект и анализ данных в реальном времени.



Они обладают **меньшими требованиями к вычислительным ресурсам** и могут быть легче интегрированы в ограниченные среды, что делает их более экономичными и доступными для различных приложений

Что Предлагают Авторы Статьи?



Простой и дешевый способ
разработки Small base LM для
последующего дообучения под
определенную задачу

Inheritune

Этапы:

1. Наследовать несколько блоков трансформеров большой LLM
2. Обучить меньшую модель на подмножестве (0,1%) необработанных данных на которых обучалась большая модель

Рассмотрены Две Ситуации

Наличие **небольшой части данных** для предварительного обучения вместе с существующей большой LM

Предобучена модель:

1.5B параметров
1B токенов данных
На 1 A6000 GPU
Меньше чем за 12 часов
На основе OpenLLmA-3B

Получение еще меньшей LM **при наличие полного датасета** для предобучения большой LM

Эксперимены с GPT2-large GPT2-medium

GPT2-large:
50% слоев, 45% параметров
GPT2-medium:
33% слоев, 28% параметров

без потери на валидационном лоссе

1. Inheritune V1

Algorithm 1 Inheritune

Require: Reference model \mathcal{M}_{ref} , a subset $\hat{\mathcal{D}}_{train}$ from \mathcal{D}_{train} used to train \mathcal{M}_{ref} , the number of layers n we want in our final model, number of epochs E , and training hyper-parameters.

- 1: Let k be the number of transformer blocks in \mathcal{M}_{ref} , with the corresponding parameters being $\theta_0, \theta_1, \dots, \theta_{k-1}$.
- 2: **Initialize:** target LM θ_{tgt} with the first n layers $\theta_0, \theta_1, \dots, \theta_{n-1}$, as well as the token embedding and prediction head layers of \mathcal{M}_{ref} .
- 3: **Train** the target LM from this initialization for E epochs using the given hyper-parameters.

Имеем:

M_{ref} – модель из k слоев

$\theta_{ref} = \{\theta_0, \theta_1, \dots, \theta_{k-1}\}$ – параметров

Обученная на D_{train} , однако в доступе лишь его часть – \hat{D}_{train}

Подмножество рандомное

Шаг 1:

Инициализация M_{tgt} посредством

наследования первых n слоев от M_{ref}

Веса M_{tgt} : $\{\theta_0, \theta_1, \dots, \theta_{n-1}\}$

Блок предсказаний и эмбединги токенов также унаследованы

Шаг 2:

Обучение M_{tgt} в течении E эпох на части данных – \hat{D}_{train}

Experimental Setup

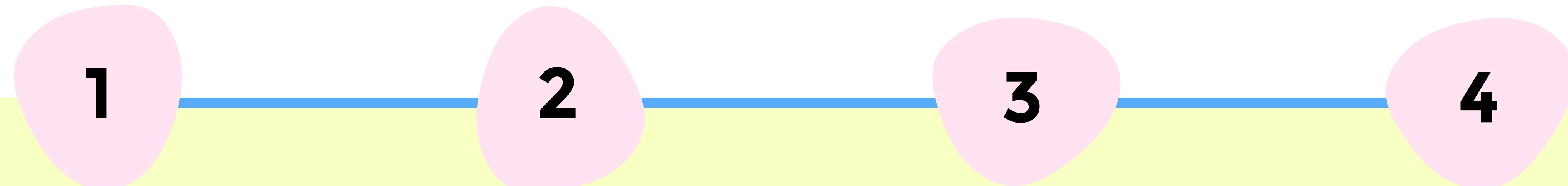
1. Данные

Исходный датасет
Redрафата v1
1 триллион токенов
википедия, книги, архивы,
stackexchange и др.

Использовано для
эксперимента
**1 миллиард
токенов**

Датасет собран в тех
же пропорциях как
предложено для
LLaMa и сделано для
OpenLLamA-3B

Итого датасет содержит
всего лишь 0.1%
данных для обучения
OpenLLamA-3B



Experimental Setup

2. Модель и Обучение

Mref – модель **OpenLLaMA-3B version1** из
k=26 слоев

Mtgt – полученная модель
n=k/2=13 слоев

E = 8 эпох (каждая эпоха использует все 1B токенов)

BatchSize = 131K токенов

Ключевым критерием при выборе исходных моделей стали данные, на которых они обучались

Models	Layers	Hidden Size	Heads
OpenLLaMA-3B	26	3200	32
OpenLLaMA-7B	32	4096	32
LLaMA2-7B	32	4096	32
GPT2-large(770M)	36	1280	20
GPT2-medium(355M)	24	1024	16

Table 1: Overview of reference models used in this study and their architectural configurations. We obtain a pre-trained OpenLLaMA-3B and we trained all GPT2 models with OpenWebText consisting of 9B tokens.

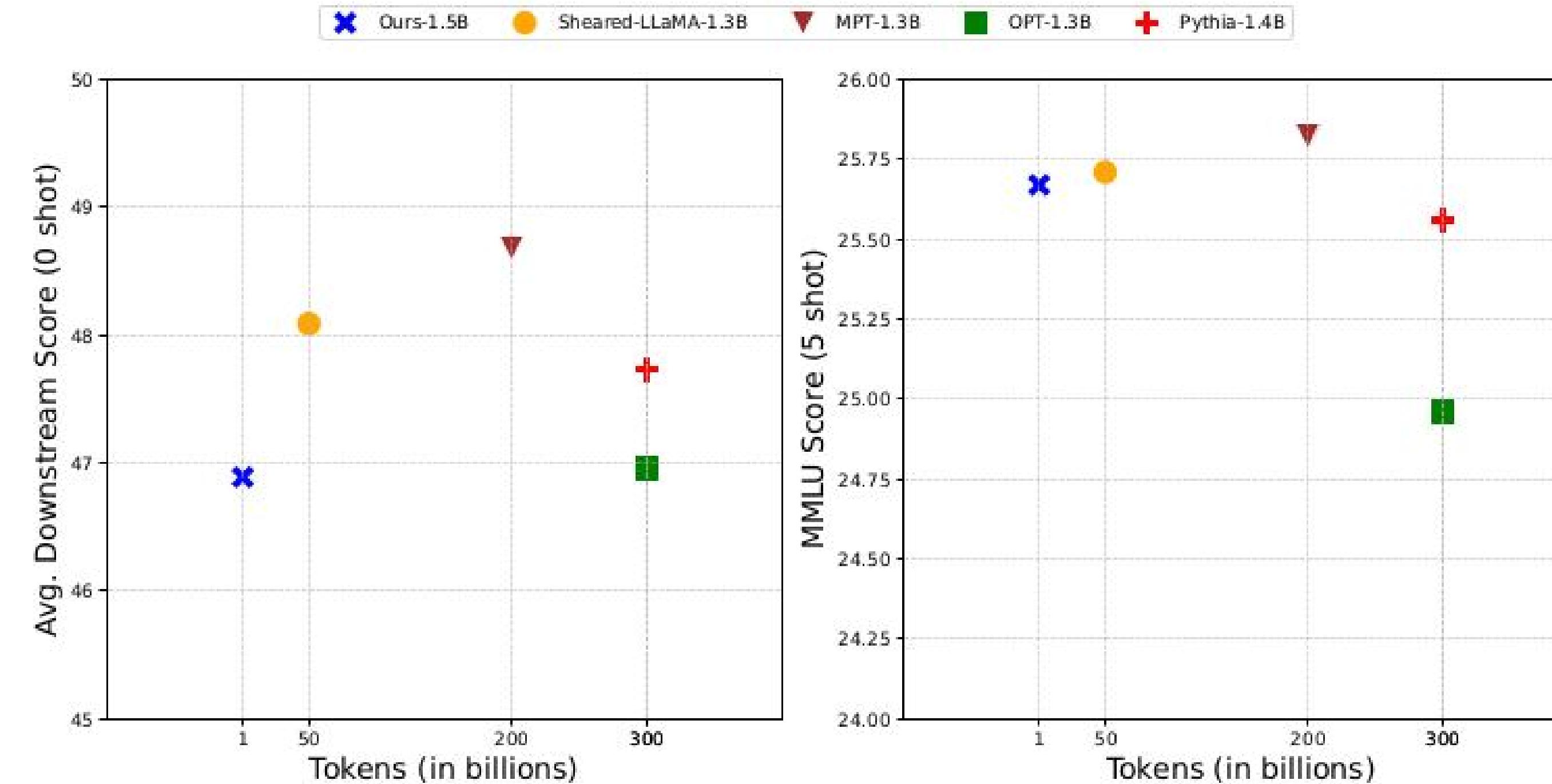
Model	Training Data (# tokens)
OpenLLaMA-3B v1(ref)	RedPajama(1T)
Ours-1.5B*	RedPajama (1B)
Shear-LLaMA-1.3B*	RedPajama(50B)
MPT-1.3B	RedPajama(200B)
Pythia-1.4B	The Pile(300B)
OPT-1.3B	Custom data(300B)

Table 2: Overview of the baseline models, pre-train data, and number of training tokens used to train these models.

Experimental Setup

3. Оценка

Средняя оценка
на 9 разных
датасетах



MMLU score

оценка здравого смысла,
правдивости, логических
выводов на
естественном языке
и понимания языка.

Вывод: модель хоть и обучена на меньшем количестве токенов
достигает результатов сравнимых с исходной большой моделью и
другими маленькими моделями

Лирическое отступление



**Методы оценки качества
работы модели в
исследовании**

LLM benchmarks = benchmarking datasets

Стандартизованные методы проверки как LLM справляется с различными видами задач

Подготовленные вопросно-ответные датасеты в разных областях, задачки по программированию и др.



Как подавать такой датасет модели?

Zero-shot: Модели дается задача без предварительных примеров или подсказок

Few-shot: Перед просьбой выполнить задачу, модели дается несколько примеров как ее выполнять

Fine-tune: Модель дообучается на данных, схожих в данными датасета на конкретную область

Как оценивать?

- **Accuracy**
- **BLEU-Score** – близость сгенерированного текста к требуемому
- **Perplexity** – насколько модель удивлена или в замешательстве



**Датасеты для оценки
здравого смысла**

0-shot

Physical Commonsense

Да/нет вопросы



To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.



a!



-
- Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]
- Q: Does France have a Prime Minister and a President?
P: ... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A: Yes. [Both are mentioned, so it can be inferred both exist.]
- Q: Have the San Jose Sharks won a Stanley Cup?
P: ... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016
A: No. [They were in the finals once, and lost.]
-

Вопрос Q, небольшой контекст P

Вопрос и два варианта ответа

Триivialно для человека - сложно для LLM

WinoGrande

Commonsense

Улучшенный WSC – собранный crowdsourcing'ом и с помощью алгоритма AfLite, чтобы очистить датасет от “подсказок” для модели, которые человек неосознанно внес в формулировку вопроса

Twin sentences			Options (answer)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because it's too <i>large</i> .	trophy / suitcase
	b	The trophy doesn't fit into the brown suitcase because it's too <i>small</i> .	trophy / suitcase
✓ (2)	a	Ann asked Mary what time the library closes, <i>because</i> she had forgotten.	Ann / Mary
	b	Ann asked Mary what time the library closes, <i>but</i> she had forgotten.	Ann / Mary
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get it <i>removed</i> .	tree / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get it <i>repaired</i> .	tree / roof
✗ (4)	a	The lions ate the zebras because they are <i>predators</i> .	lions / zebras
	b	The lions ate the zebras because they are <i>meaty</i> .	lions / zebras

Figure 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word “predators” is more often associated with the word “lions”, compared to “zebras”

Примеры проблем с WSC – например в последнем вопросе упомянуто слово “хищник”, которое и так ближе ко львам, чем к зебрам

LogiQA

Задачки на логику

P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

Q: What can be inferred from the above statements?

- A. Jack gets more salary than Mark.
- B. David gets the same salary as Mark.
- C. One employee supervises another who gets more salary than himself.
- ✓ D. One employee supervises another who gets less salary than himself.

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: What may be the missing premise of the above argument?

- A. Some textile workers do not work in workshop B.
- B. Some employees working in workshop B are not textile workers.
- ✓ C. Some textile workers work in workshop B.
- D. Some employees living in dormitory area A work in the workshop B.

Figure 1: Examples of LogiQA. (✓ indicates the correct answer.)



**Датасеты для оценки
понимания и
воспроизведения
естественного языка**

0-shot

MNLI

The Multi-Genre Natural Language
Inference Corpus

Даются высказывание и гипотеза

Модель должна понять гипотеза

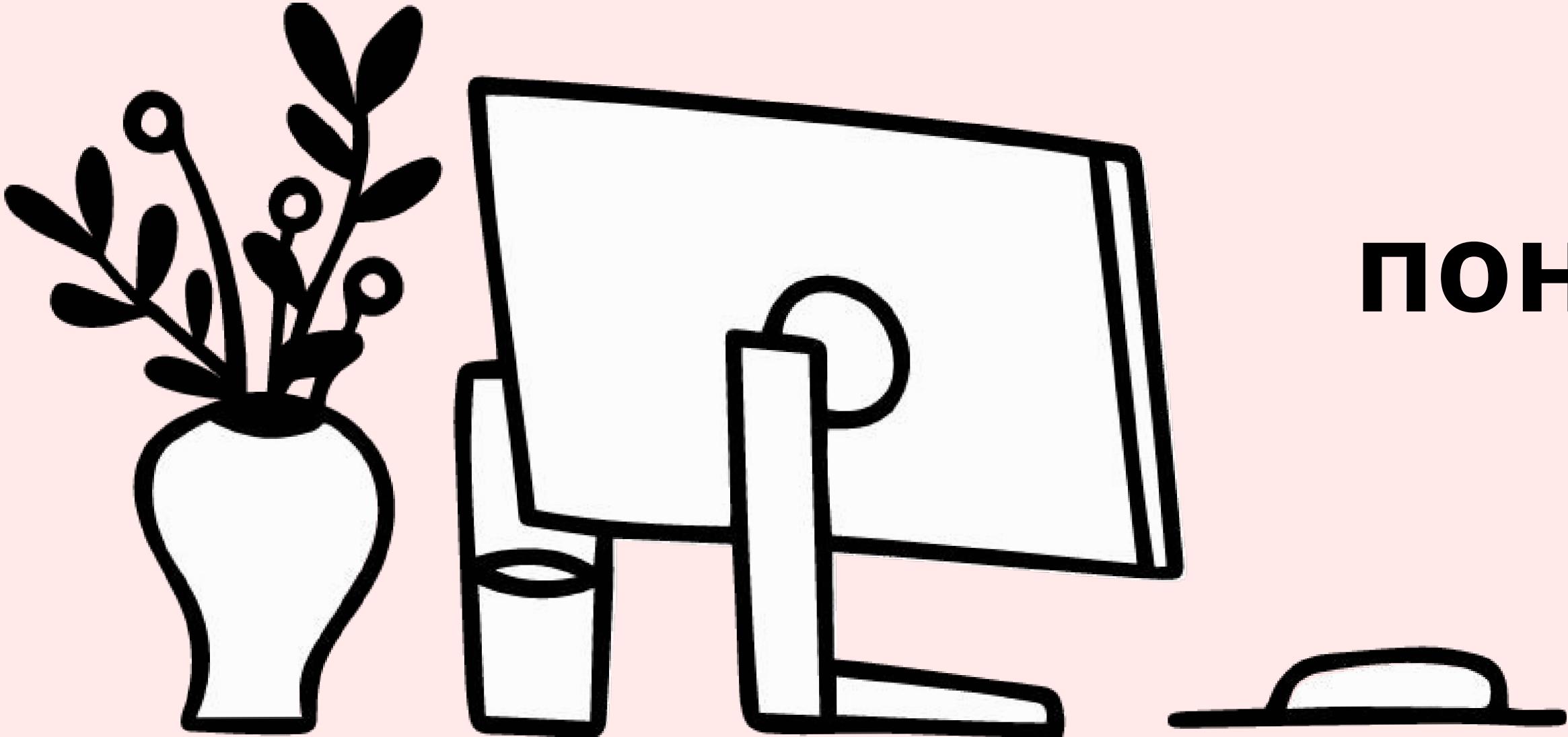
- а) сопровождает высказывание
- б) противоречит высказыванию
- в) нейтрально

QNLI

Вопросно-ответный датасет, состоящий из пар "вопрос-абзац", где одно из предложений в абзаце (взятое из Википедии) содержит ответ на соответствующий вопрос

WNLI

Основан на упомянутой ранее WSC –
Winograd Schema Challenge



Датасет на общее понимание естественного языка

5-shot

MMLU (5-shot)

Find all c in \mathbb{Z}_3 such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.
(A) 0 (B) 1 (C) 2 (D) 3

Figure 14: An Abstract Algebra example.

What is the embryological origin of the hyoid bone?
(A) The first pharyngeal arch
(B) The first and second pharyngeal arches
(C) The second pharyngeal arch
(D) The second and third pharyngeal arches

Figure 15: An Anatomy example.

Why isn't there a planet where the asteroid belt is located?
(A) A planet once formed here but it was broken apart by a catastrophic collision.
(B) There was not enough material in this part of the solar nebula to form a planet.
(C) There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.
(D) Resonance with Jupiter prevented material from collecting together to form a planet.

Figure 16: An Astronomy example.

Three contrasting tactics that CSO's can engage in to meet their aims are _____ which typically involves research and communication, _____, which may involve physically attacking a company's operations or _____, often involving some form of _____.
(A) Non-violent direct action, Violent direct action, Indirect action, Boycott
(B) Indirect action, Instrumental action, Non-violent direct action, Information campaign
(C) Indirect action, Violent direct action, Non-violent direct-action Boycott.
(D) Non-violent direct action, Instrumental action, Indirect action, Information campaign

Figure 17: A Business Ethics example.

● Примеры из разных областей
(57 видов задач)

● 5-shot: перед оценкой на датасете дается модели пять примеров для каждой задачи

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$

Answer: A

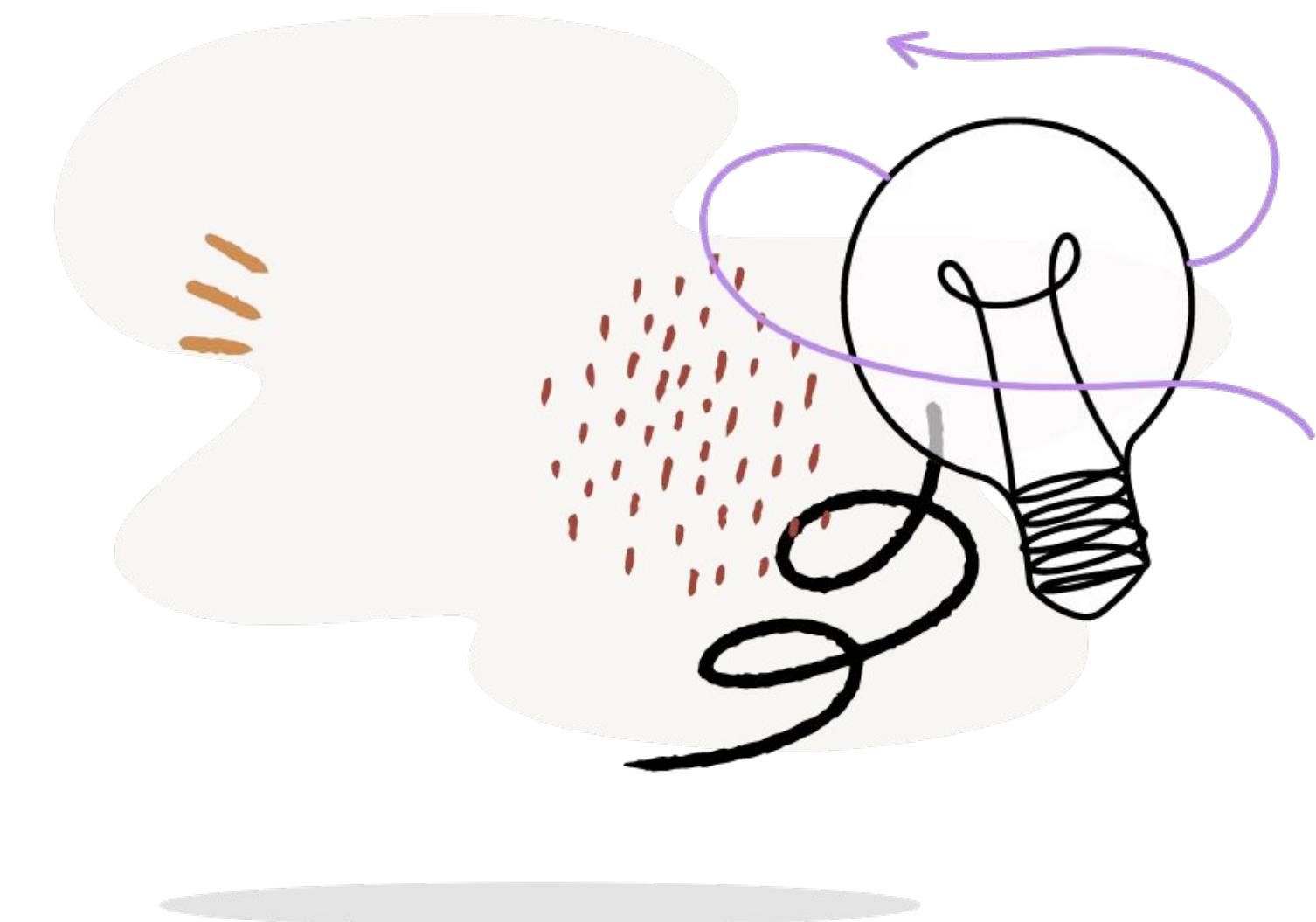
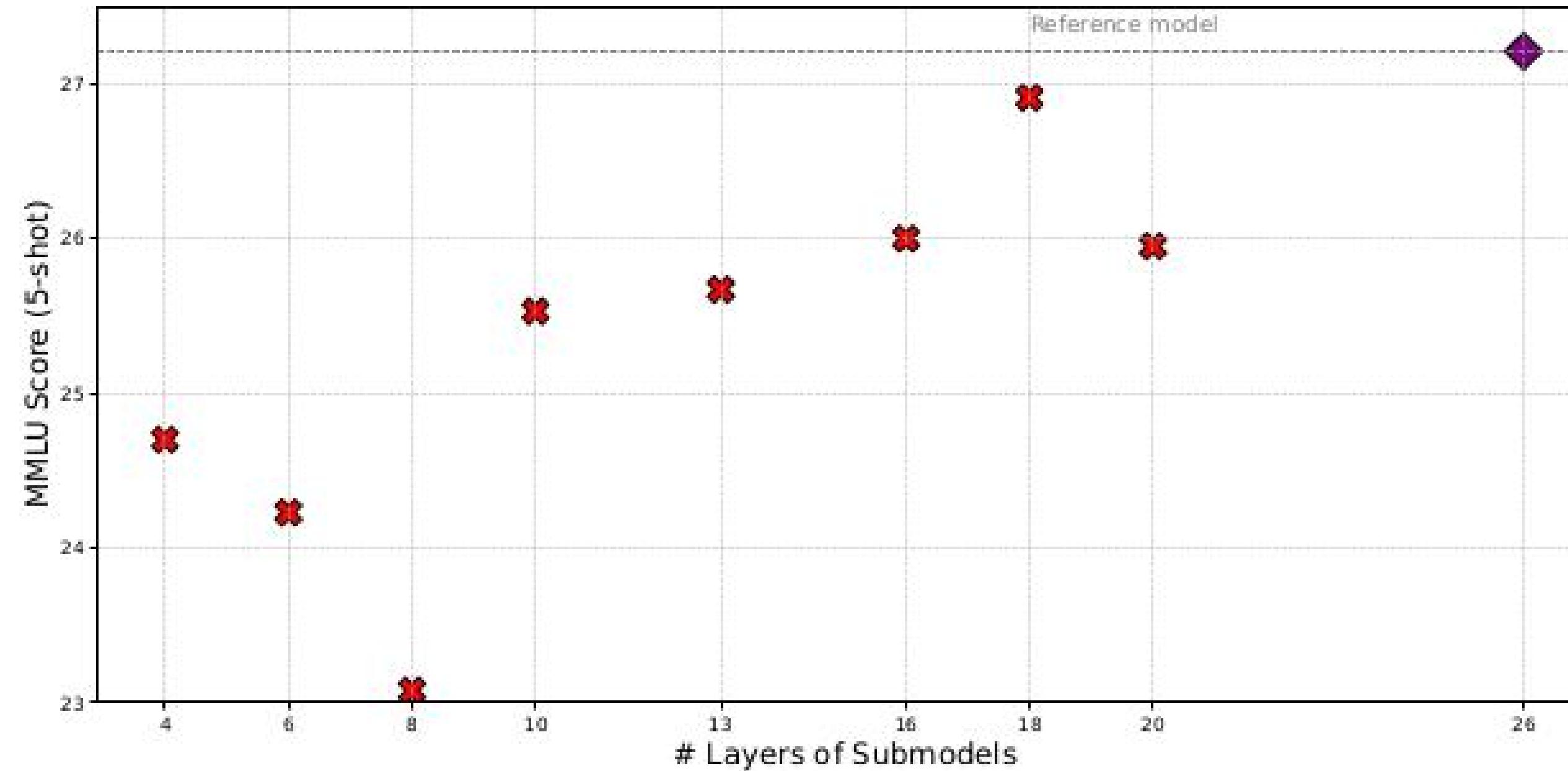
If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30

Answer: C

2-shot примеры

C – ответ модели

Дополнительные Выводы



Метод масштабируется на модели разного размера (с разным количеством слоев)

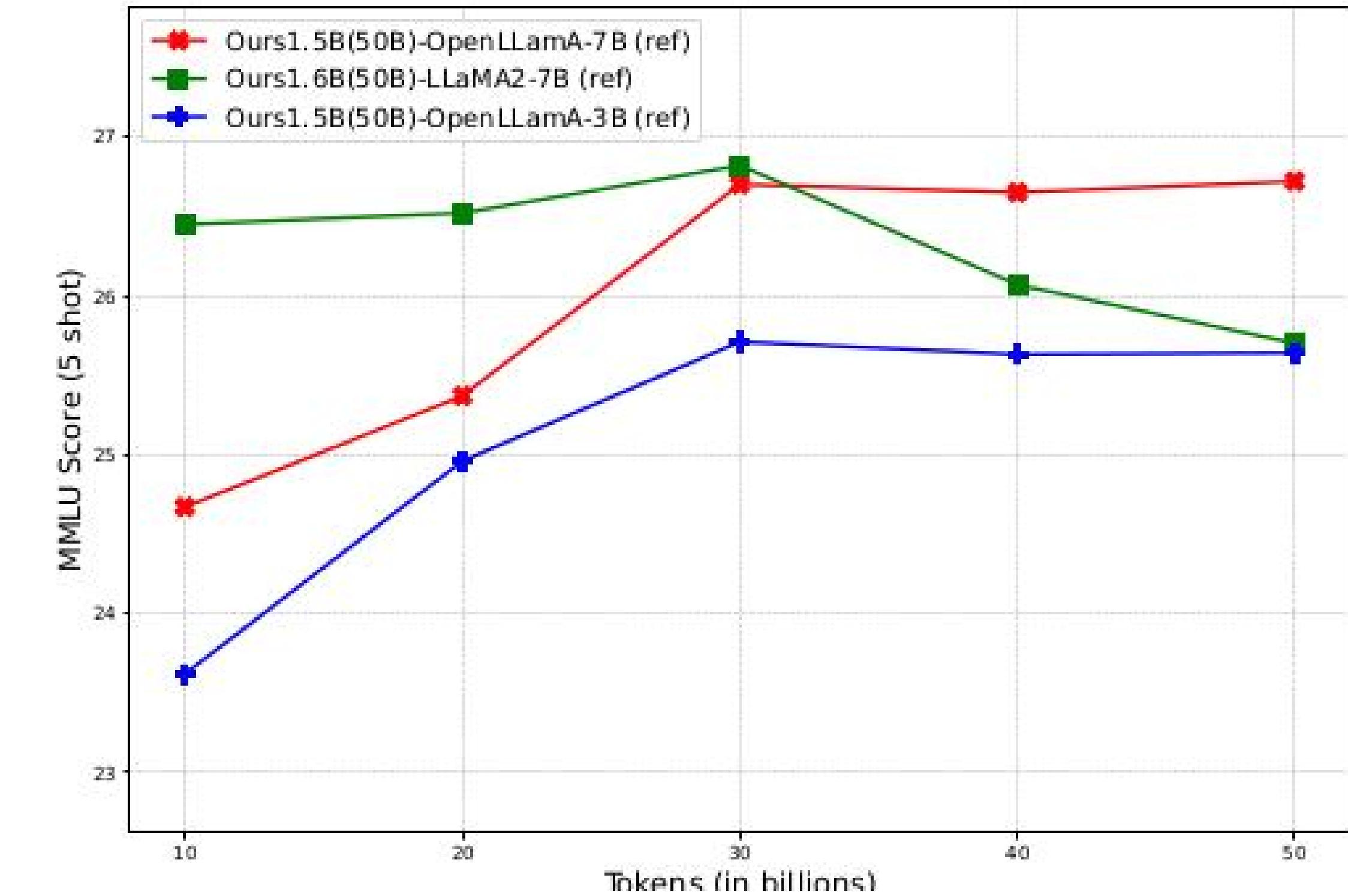
Оценка:
MMLU (5-shot) benchmark

Дополнительные Выводы

Анализ с исходной моделью большего размера и 50B токенов данных

Model (# tokens), ref	MMLU(5)
Ours-1.6B (1B), LLaMA2-7B	24.27
Ours-1.5B (1B), OpenLLaMA-3B	25.67
Ours-1.5B (50B), OpenLLaMA-3B	25.71
Ours-1.6B (50B), LLaMA2-7B	26.07
Ours-1.6B (50B), OpenLLaMA-7B	26.72

Даже обученные на меньшем количестве токенов модели показывают конкурентные результаты



Явные улучшения с увеличением количества данных

Дополнительные Выводы

**Следует ли повторно
использовать токены 1B в течение нескольких эпох или использовать то же
количество новых токенов?**

Model (# tokens)	Data type	MMLU (5-shot)
Ours-1.5B (1B)	10 epochs	24.95
Ours-1.5B (50B)	10B fresh	23.62
Ours-1.5B (1B)	20 epochs	25.46
Ours-1.5B (50B)	20B fresh	24.96

**Можно безопасно повторно использовать
1B токенов до 10–20 эпох**

4. Сравнение С Другими Моделями

Experimental Setup

Model		Commonsense Reasoning				
Name (# train tokens)	Reference	Winograd	PIQA	Boolq	WinoGrande	Logiqa
OpenLLaMA-3B (1T)	n/a	63.46	74.97	67.18	62.27	28.4
OPT-1.3B (300B)	n/a	38.46	71.82	57.83	59.51	27.04
Pythia-1.4B (300B)	n/a	36.54	70.89	63.12	56.99	27.65
MPT-1.3B (200B)	n/a	63.46	71.44	50.89	58.09	28.26
Sheared LLaMA-1.3B (50B)	LLaMA2-7B	36.54	73.45	62.02	58.17	27.34
Ours-1.5B (1B)	OpenLLaMA-3B	50.96	56.47	61.68	51.69	25.19

Model		Lang. Understanding & Inference			Factuality	
Name (# train tokens)	Reference	MMLU(5)	WNLI	QNLI	MNLI	TruthfulQA
OpenLLaMA-3B (1T)	n/a	27.21	50.7	51.3	37.3	35
OPT-1.3B (300B)	n/a	24.96	42.25	51.29	35.82	38.67
Pythia-1.4B (300B)	n/a	25.56	53.52	49.48	32.76	38.66
MPT-1.3B (200B)	n/a	25.82	40.85	50.52	35.93	38.68
Sheared LLaMA-1.3B (50B)	LLaMA2-7B	25.71	49.3	50.98	37.94	37.14
Ours-1.5B (1B)	OpenLLaMA-3B	25.67	43.66	49.41	34.42	48.61

Сравнение с моделями схожего размера, но обученных с нуля
А также с моделью ShearedLLaMa обученной в технике pruning

Выделены скоры, если модель достигла 90% качества исходной модели, либо скор лучше чем как минимум у двух других

Лирическое отступление



**Pruning technique –
избавление от
избыточных слоев сети
для ускорения inference
без потери точности**

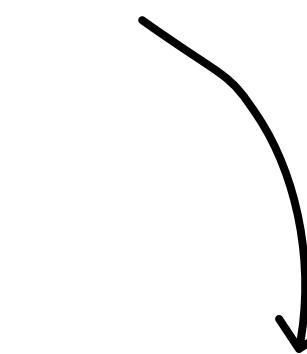
[Статья](#)

Pruning - одна из техник сжатия моделей

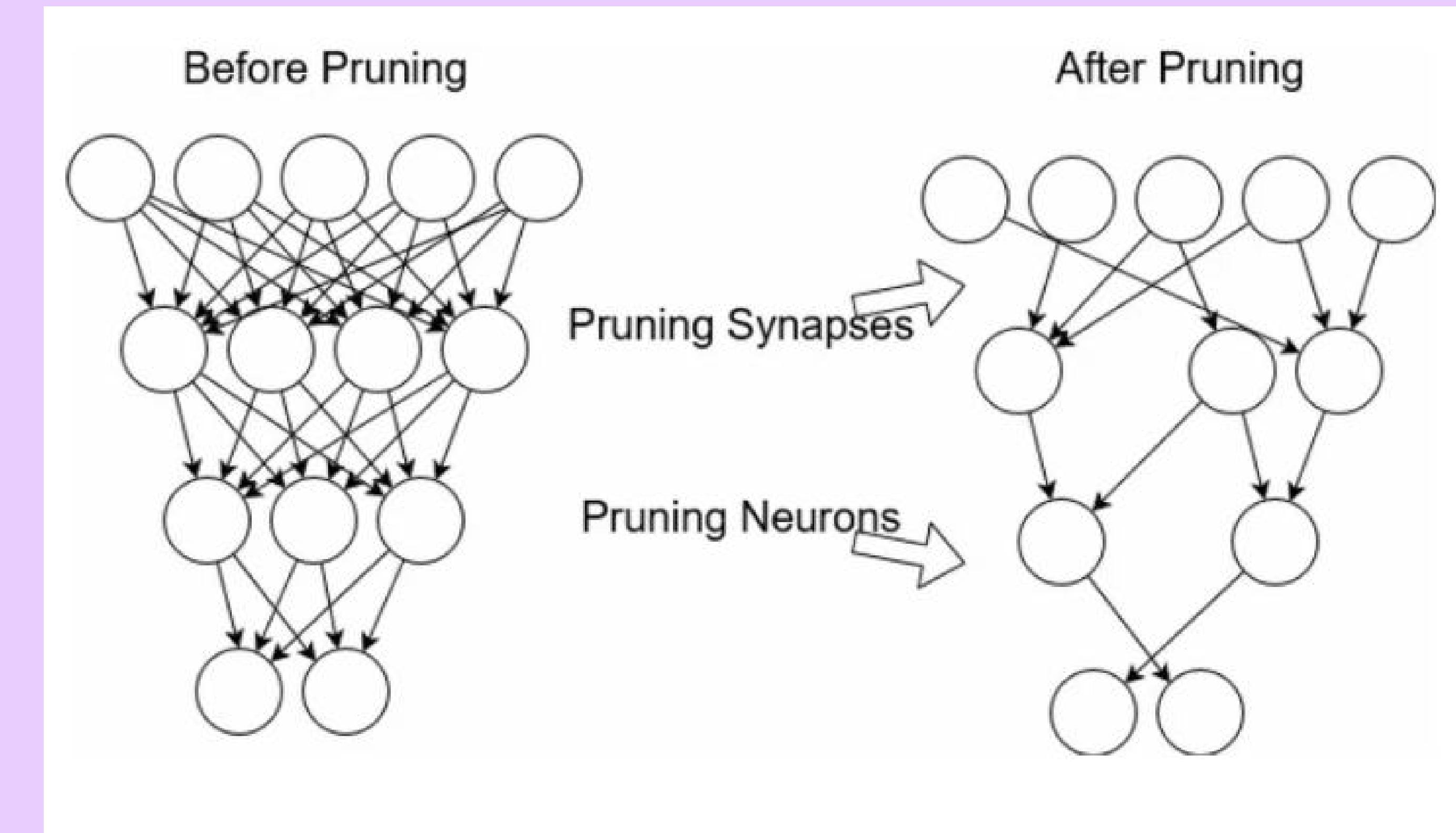
Популярные техники

- **pruning**
- quantization
- knowledge distillation
- low-rank factorization

Однако для LLM все-равно урезается производительность



Авторы статьи предлагают способ, показывающий лучшие результаты, чем просто pruning



Урезание весов/нейронов/слоев и так далее

Алгоритм создания

Sheared-LLaMA-1.3B, Sheared-LLaMA-2.7B

1. Targeted structured pruning

Урезание до определенной target архитектуры

2. Dynamic batch loading

Подгрузка данных из
каждой
предметной области
пропорционально
уменьшению loss'a

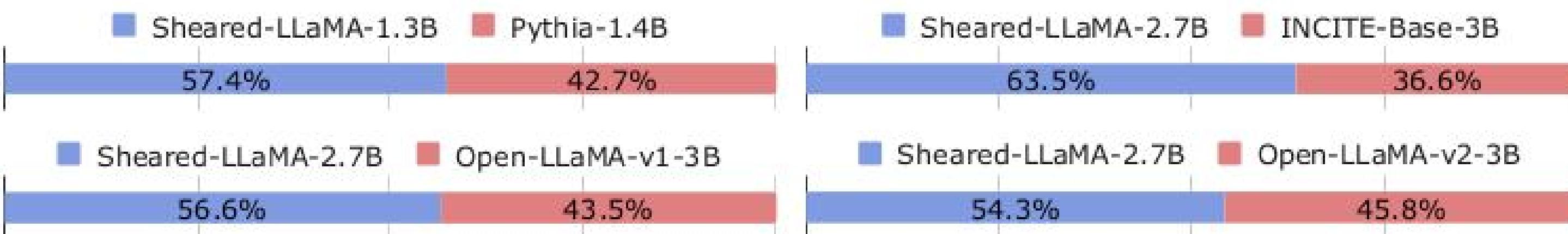
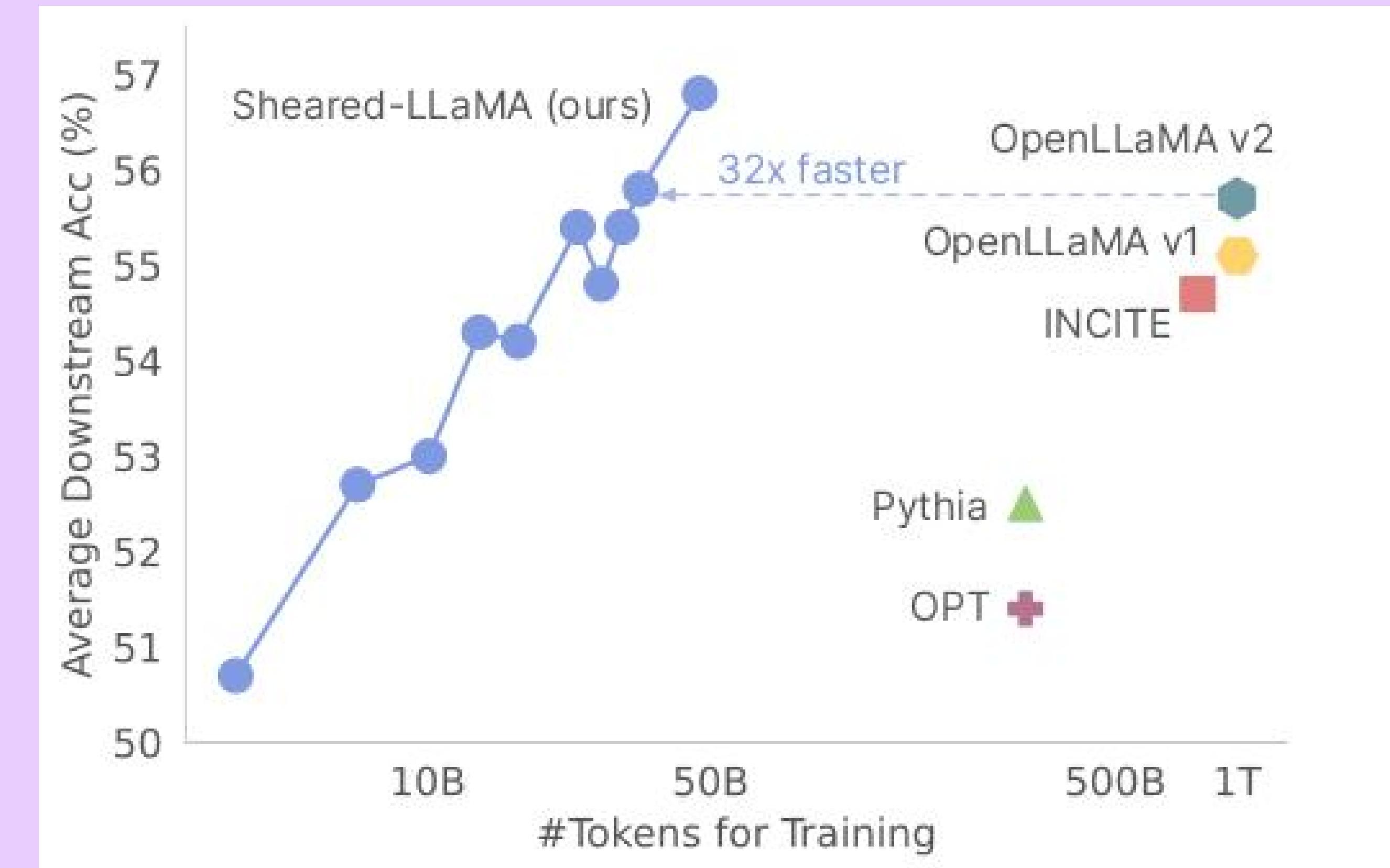


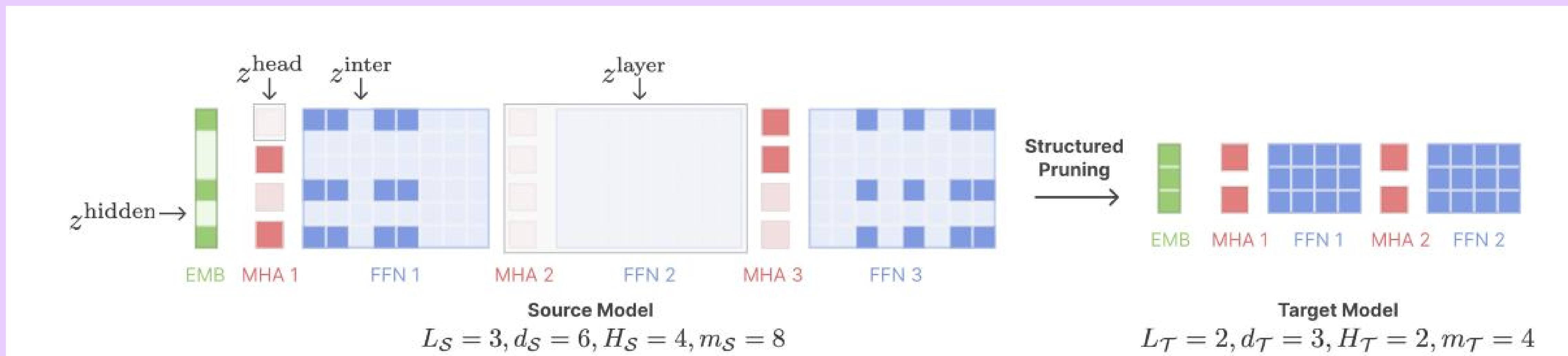
Figure 3: Sheared-LLaMAs outperform Pythia-1.4B, INCITE-Base-3B, OpenLLaMA-3B-v1 and OpenLLaMA-3B-v2 in instruction tuning.

Сравнение с другими моделями

Targeted structured pruning

Набор корректирующих масок для параметров модели на разных уровнях

Granularity	Layer	Hidden dimension	Head	Intermediate dimension
Pruning masks	$z^{\text{layer}} \in \mathbb{R}^{L_S}$	$z^{\text{hidden}} \in \mathbb{R}^{d_S}$	$z^{\text{head}} \in \mathbb{R}^{H_S} (\times L_S)$	$z^{\text{int}} \in \mathbb{R}^{m_S} (\times L_S)$



$z = 0$ – удаляем , $z = 1$ оставляем соответствующие параметры

Обучение масок – задача оптимизации

Dynamic Batch Loading

Algorithm 1: Dynamic Batch Loading

Require: Training data of k domains D_1, D_2, \dots, D_k , validation data $D_1^{\text{val}}, D_2^{\text{val}}, \dots, D_k^{\text{val}}$, initial data loading weights $w_0 \in \mathbb{R}^k$, reference loss $\ell_{\text{ref}} \in \mathbb{R}^k$, LM loss \mathcal{L} or pruning loss $\mathcal{L}_{\text{prune}}$, training steps T , evaluation per m steps, model parameters θ (θ, z, ϕ, λ for pruning)

```
for  $t = 1, \dots, T$  do
    if  $t \bmod m = 0$  then
         $\ell_t[i] \leftarrow \mathcal{L}(\theta, z, D_i^{\text{val}})$  if pruning else  $\mathcal{L}(\theta, D_i^{\text{val}})$ 
         $\Delta_t[i] \leftarrow \max \{\ell_t[i] - \ell_{\text{ref}}[i], 0\}$                                 ▷ Calculate loss difference
         $w_t \leftarrow \text{UpdateWeight}(w_{t-m}, \Delta_t)$                                 ▷ Update data loading proportion
    end
```

Sample a batch of data \mathcal{B} from D_1, D_2, \dots, D_k with proportion w_t ;

```
if pruning then
    Update  $\theta, z, \phi, \lambda$  with  $\mathcal{L}_{\text{prune}}(\theta, z, \phi, \lambda)$  on  $\mathcal{B}$ 
else
    Update  $\theta$  with  $\mathcal{L}(\theta, \mathcal{B})$ 
end
```

end

Subroutine $\text{UpdateWeight}(w, \Delta)$

```
 $\alpha \leftarrow w \cdot \exp(\Delta)$                                 ▷ Calculate the unnormalized weights
 $w \leftarrow \frac{\alpha}{\sum_i \alpha[i]}$  return  $w$                                 ▷ Renormalize the data loading proportion
return  $\theta$ 
```

На каждом шаге подается
для каждой группы данных
 D соответствующая доля
данных $w[t]$

И далее обновляются
параметры модели θ на батче
этих данных

Алгоритм применялся и для
pruning этапа и для
 дальнейшего pre-training
этапа

Dynamic Batch Loading

Algorithm 1: Dynamic Batch Loading

Require: Training data of k domains D_1, D_2, \dots, D_k , validation data $D_1^{\text{val}}, D_2^{\text{val}}, \dots, D_k^{\text{val}}$, initial data loading weights $w_0 \in \mathbb{R}^k$, reference loss $\ell_{\text{ref}} \in \mathbb{R}^k$, LM loss \mathcal{L} or pruning loss $\mathcal{L}_{\text{prune}}$, training steps T , evaluation per m steps, model parameters θ (θ, z, ϕ, λ for pruning)

for $t = 1, \dots, T$ **do**

if $t \bmod m = 0$ **then**

$\ell_t[i] \leftarrow \mathcal{L}(\theta, z, D_i^{\text{val}})$ if *pruning* else $\mathcal{L}(\theta, D_i^{\text{val}})$
 $\Delta_t[i] \leftarrow \max \{\ell_t[i] - \ell_{\text{ref}}[i], 0\}$
 $w_t \leftarrow \text{UpdateWeight}(w_{t-m}, \Delta_t)$

 ▷ Calculate loss difference
 ▷ Update data loading proportion

end

 Sample a batch of data \mathcal{B} from D_1, D_2, \dots, D_k with proportion w_t ;

if *pruning* **then**

 Update θ, z, ϕ, λ with $\mathcal{L}_{\text{prune}}(\theta, z, \phi, \lambda)$ on \mathcal{B}

else

 Update θ with $\mathcal{L}(\theta, \mathcal{B})$

end

end

Subroutine $\text{UpdateWeight}(w, \Delta)$

$\alpha \leftarrow w \cdot \exp(\Delta)$
 $w \leftarrow \frac{\alpha}{\sum_i \alpha[i]}$ **return** w

return θ

Каждые m шагов получали loss текущей модели на валидации

Δ – Разницу между текущим loss'ом и reference loss

Обновление доли $w[t]$ на основе этой разницы

Обновление доли $w[t]$ происходило экспоненциально

4. Сравнение С Другими Моделями

Experimental Setup

Model		Commonsense Reasoning				
Name (# train tokens)	Reference	Winograd	PIQA	Boolq	WinoGrande	Logiqa
OpenLLaMA-3B (1T)	n/a	63.46	74.97	67.18	62.27	28.4
OPT-1.3B (300B)	n/a	38.46	71.82	57.83	59.51	27.04
Pythia-1.4B (300B)	n/a	36.54	70.89	63.12	56.99	27.65
MPT-1.3B (200B)	n/a	63.46	71.44	50.89	58.09	28.26
Sheared LLaMA-1.3B (50B)	LLaMA2-7B	36.54	73.45	62.02	58.17	27.34
Ours-1.5B (1B)	OpenLLaMA-3B	50.96	56.47	61.68	51.69	25.19

Model		Lang. Understanding & Inference			Factuality	
Name (# train tokens)	Reference	MMLU(5)	WNLI	QNLI	MNLI	TruthfulQA
OpenLLaMA-3B (1T)	n/a	27.21	50.7	51.3	37.3	35
OPT-1.3B (300B)	n/a	24.96	42.25	51.29	35.82	38.67
Pythia-1.4B (300B)	n/a	25.56	53.52	49.48	32.76	38.66
MPT-1.3B (200B)	n/a	25.82	40.85	50.52	35.93	38.68
Sheared LLaMA-1.3B (50B)	LLaMA2-7B	25.71	49.3	50.98	37.94	37.14
Ours-1.5B (1B)	OpenLLaMA-3B	25.67	43.66	49.41	34.42	48.61

Таким образом, сравнение с ShearedLLaMa не совсем корректное, так как

ShearedLLaMa создана с помощью **target structural pruning** на 0,4B токенах и после предобучена на 50B токенах

Inheritune – это больше техника для лучшей инициализации модели

Выводы:

1. Быстрое предобучение

1.5B модель на 1 A6000 GPU за 12 часов

Для сравнения:

- mpt-1.3B 440 A100 GPUs полдня
- Pythia-1.4B 64 A100-40GB GPUs for 4830 GPU hours
- TinyLLaMA-1.1B 16 A100 GPUs 3 месяца

Простой и дешевый способ
разработки Small base LM для
последующего дообучения под
определенную задачу

2. Для Small LM достаточно всего лишь части данных и нескольких слоев Large LM

Inheritune

Ограничения:

1. Можно изменять только количество блоков трансформера/слоев

Что ограничивает возможности изменения архитектуры **Small LM**, например, скрытых слоев и голов внимания

2. Метод скорее всего сильно чувствителен к датасету, так как данных используется мало
3. В работе нет выводов относительно того, какие лучше блоки модели выбирать и какой датасет

Простой и дешевый способ разработки **Small base LM** для последующего дообучения под определенную задачу

Inheritune

**Благодарю за
внимание**



Ссылки приведенные в презентации

Статья Inheritune

Репозиторий Inheritune

PIQA benchmark dataset

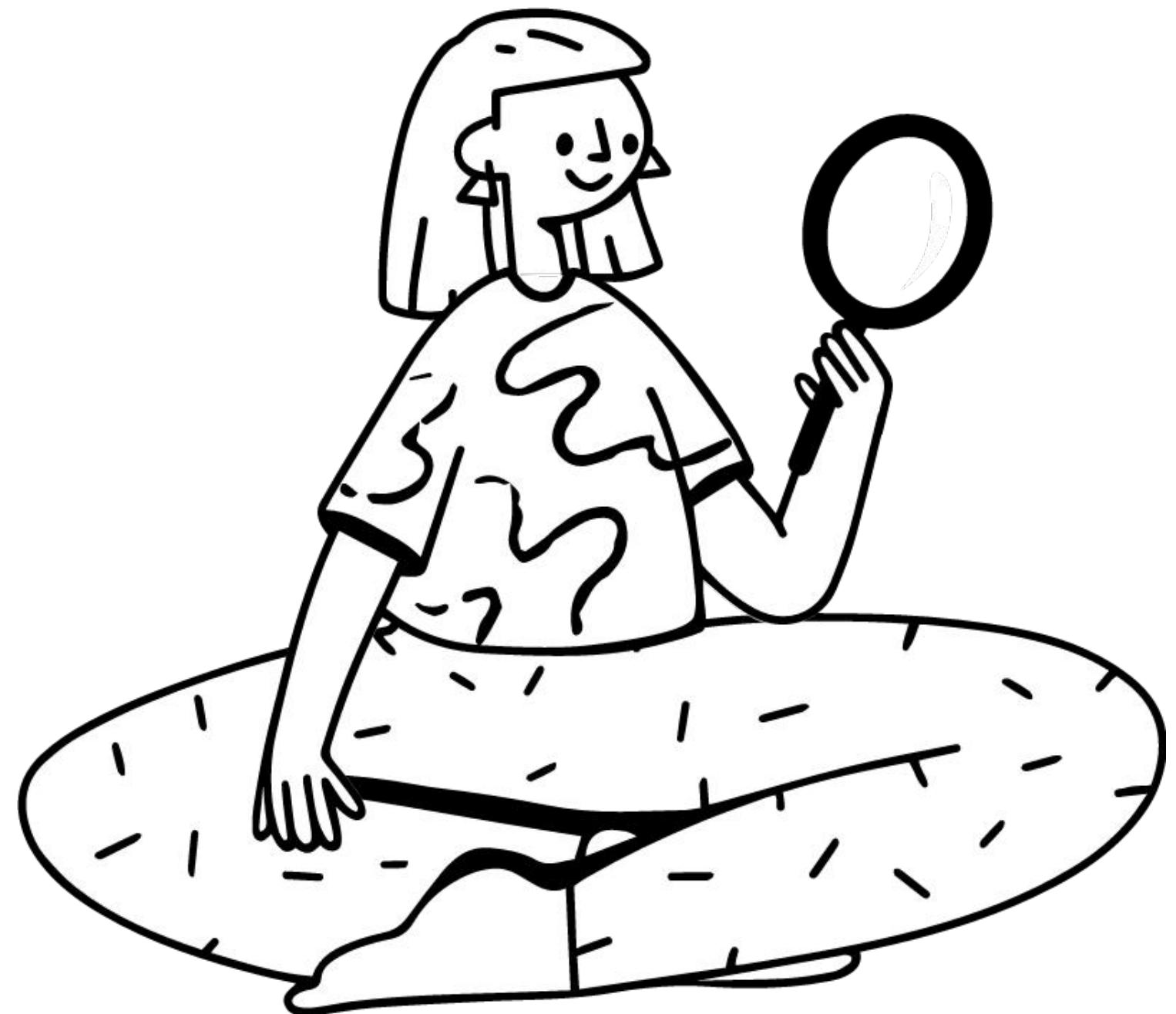
BoolQ benchmark dataset

WinoGrande benchmark dataset

LogiQA benchmark dataset

MNLI QNLI WNLI benchmark datasets

MMLU benchmark dataset



Статья Sheared-LLaMa structured pruning