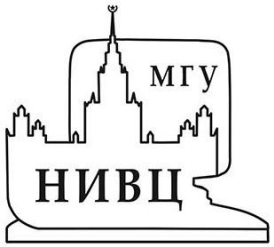


Данные

к.ф.-м.н. Тихомиров М.М.

НИВЦ МГУ имени М. В. Ломоносова

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

Hoffmann J. et al. Training compute-optimal large language models. arXiv //arXiv preprint arXiv:2203.15556. – 2022.

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

	Approach 2		Approach 3	
Parameters	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

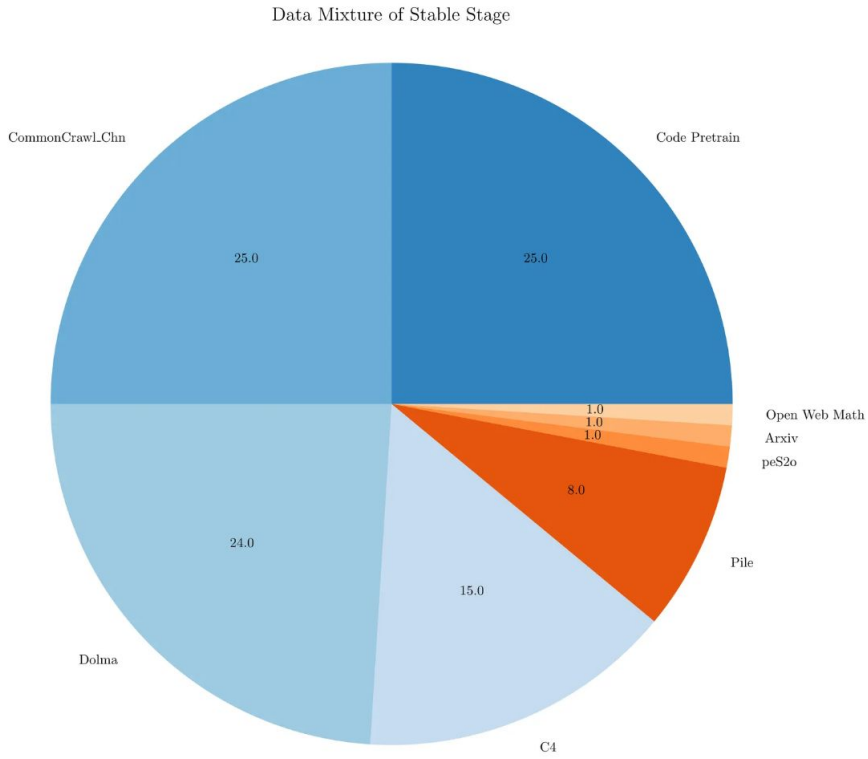
Gopher - 280B, 300B
Chinchilla - 70B, 1.4T

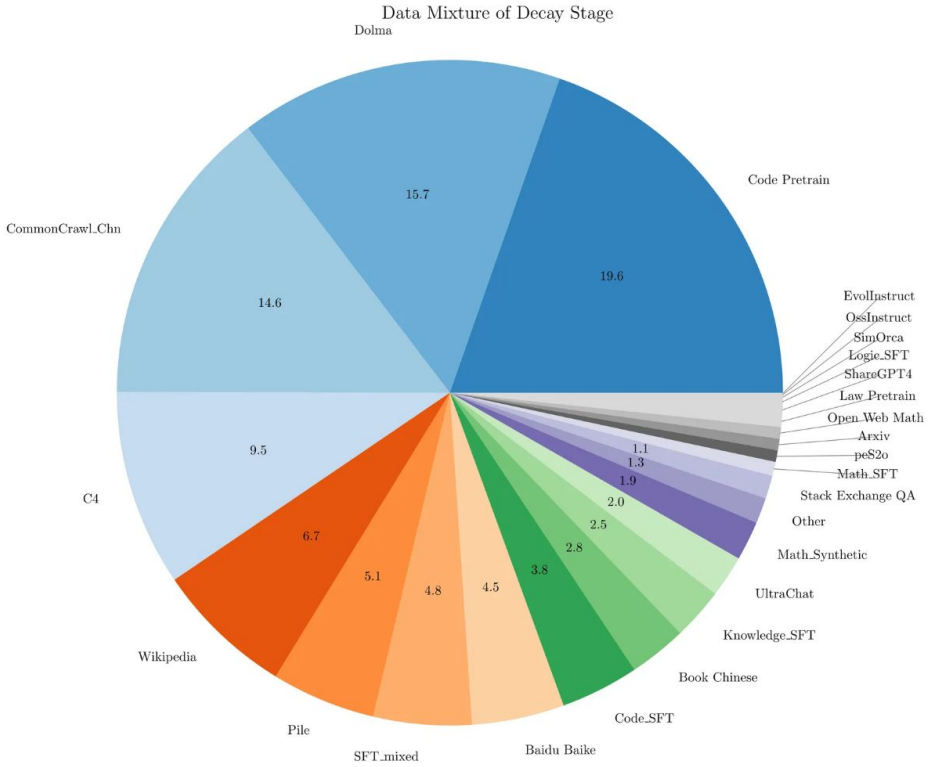
Итоговый компьют (flops) одинаковый!

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_macroeconomics	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			

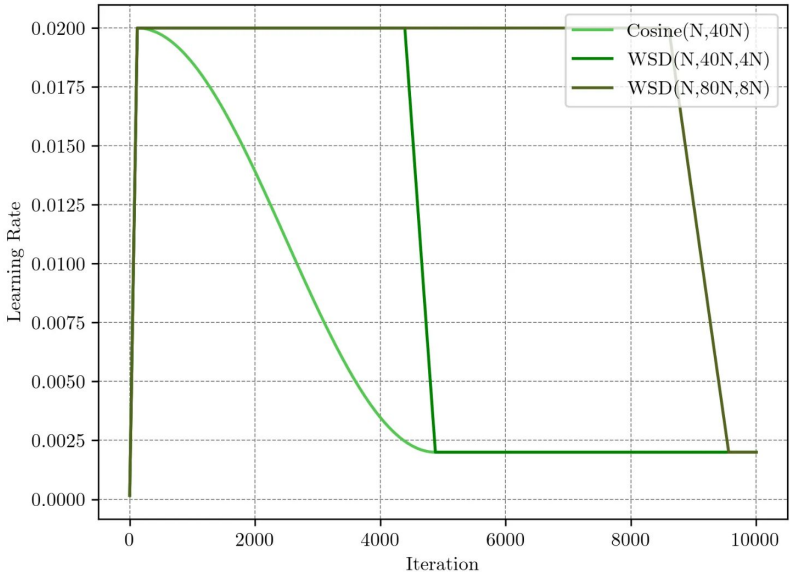
Table A6 | ***Chinchilla* MMLU results.** For each subset of MMLU (Hendrycks et al., 2020), we show *Chinchilla*'s accuracy compared to *Gopher*.

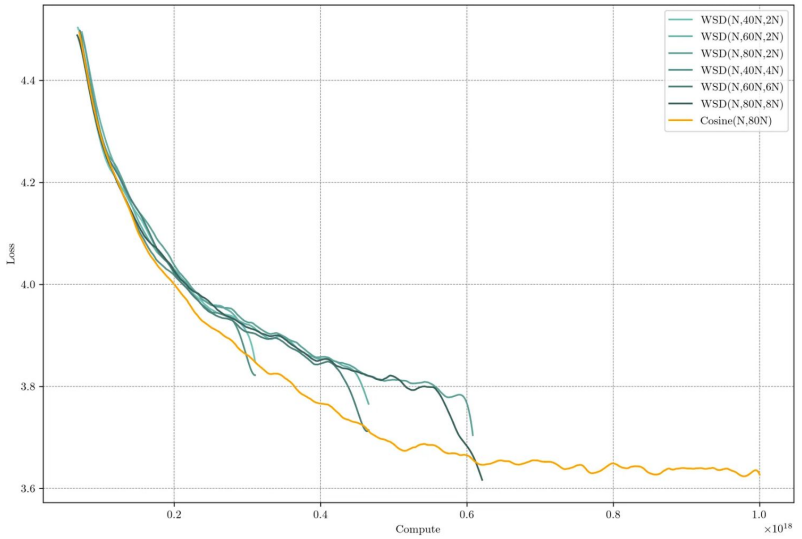
MiniCPM подход: два этапа обучения





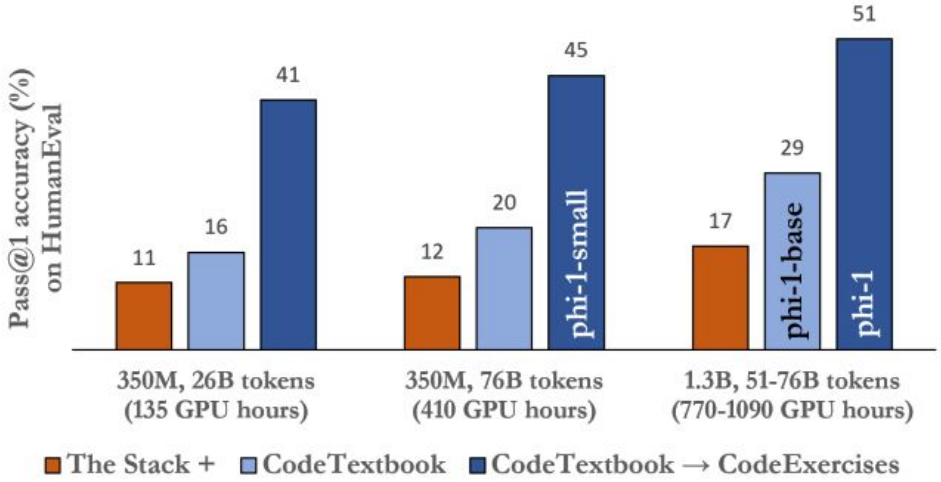
MiniCPM подход: annealing phase





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

Textbooks Are All You Need (Phi-1)



The Stack - filtered code-language dataset (6B токенов)

CodeTextbook - synthetic textbook (1B т., генерировали через GPT-3.5)

Code Exercises - также синтетика, но instruct (~180M токенов)

Phi-3: развитие идеи Phi-1

Наращивание масштабов

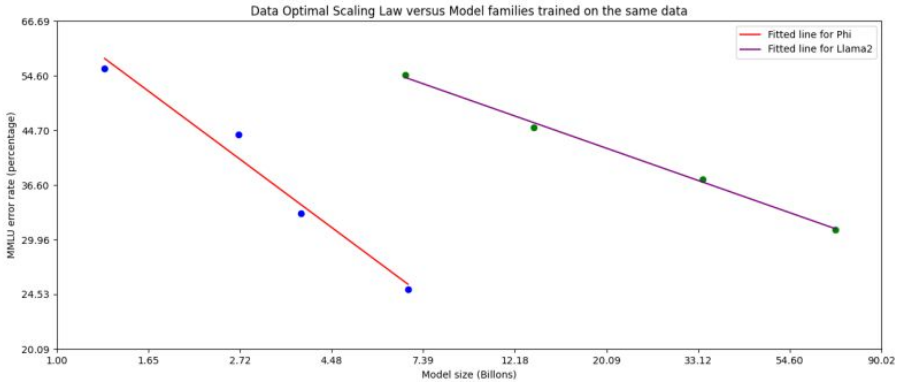
4.5T токенов

Смешивание синтетики + Web дата (качественных)

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21a]	68.8	75.7	78.0	56.3	61.7	63.6	66.5	70.5	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7	77.0	82.4	53.6	58.5	49.8	71.1	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8	58.1	55.8	42.5	47.1	48.7	57.3	55.2	58.1
GSM-8K (8-Shot; CoT) [CKB+21]	82.5	89.6	91.0	61.1	46.4	59.8	77.4	64.7	78.1
MATH (0-Shot; CoT) [HBK+21b]	41.3	34.6	53.1	–	15.0	13.6	28.2	11.1	45.3
MedQA (2-Shot) [JPO+20]	53.8	65.4	69.9	40.9	50.0	49.6	60.5	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5	45.1	50.2	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	58.1	73.9	45.2	75.2	72.3	67.7	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9	90.7	91.6	75.9	78.6	78.3	82.8	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6	97.0	97.7	88.5	90.6	91.4	93.4	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	86.9	87.9	60.2	77.7	78.1	75.7	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.2	80.2	68.3	74.6	65.5	73.9	75.9	68.3
BigBench-Hard (3-Shot; CoT) [SRR+22, SSS+22]	71.7	79.1	81.4	59.4	57.3	59.6	51.5	69.7	68.32
WinoGrande (5-Shot) [SLBBC19]	70.8	81.5	81.5	54.7	54.2	55.6	65.0	62.0	68.8
OpenBookQA (10-Shot) [MCKS18]	83.2	88.0	87.4	73.6	79.8	78.6	82.6	85.8	86.0
BoolQ (2-Shot) [CLC+19]	77.2	84.8	86.5	–	72.2	66.0	80.9	77.6	79.1
CommonSenseQA (10-Shot) [THLB19]	80.2	80.0	82.8	69.3	72.6	76.2	79.0	78.1	79.6
TruthfulQA (10-Shot; MC2) [LHE22]	65.0	70.2	75.1	–	53.0	52.1	63.2	60.1	85.8
HumanEval (0-Shot) [CTJ+21]	58.5	61.0	62.2	59.0	28.0	34.1	60.4	37.8	62.2
MBPP (3-Shot) [AON+21]	70.0	71.7	75.2	60.6	50.8	51.5	67.7	60.2	77.8
Average	69.7	73.6	76.7	–	58.9	59.3	67.3	66.8	72.8
GPQA (2-Shot; CoT) [RHS+23]	32.8	34.3	–	–	–	–	–	–	29.0
MT Bench (2 round ave.) [ZCS+23]	8.38	8.70	8.91	–	–	–	–	–	8.35

Phi-1-2-3 некоторые выводы

- Первые две версии обвиняли в train on test
- 3 версия вышла хорошей
  - **По метрикам** обходит всех в своей “весовой”
- Секрет успеха: комбинация синтетики и настоящих данных, но в итоге 3 версия обучалась на 4.5T токенах
- Качество данных **существенно!**



Источники данных

Откуда берут данные?

- Wikipedia
- Новости
- Reddit / Pikabu / иные платформы
- Книги, учебники
- Web (основной)
- GitHub
- StackOverflow
- Видео? OCR?

Common Crawl

- 20 TB каждый месяц
- “Грязные” данные
- Настоящих текстов в реальности сильно меньше
- Основа большинства датасетов

Distribution of Languages

The language of a document is identified by Compact Language Detector 2 (CLD2). It is able to identify 160 different languages and up to 3 languages per document. The table lists the percentage covered by the primary language of a document (returned first by CLD2). So far, only HTML pages are passed to the language detector. The underlying data including page counts is provided in languages.csv.

crawl	CC-MAIN-2024-33	CC-MAIN-2024-38	CC-MAIN-2024-42
language	%	%	%
eng	43.1787	44.1210	43.4241
rus	6.2242	6.1556	6.0444
zho	5.1917	4.6266	4.8129
deu	5.0895	5.4471	5.3038
jpn	4.8790	5.1119	5.0419
spa	4.5819	4.4769	4.5387
fra	4.2439	4.4292	4.3960
<unknown>	3.2930	2.6706	3.2780
ita	2.4921	2.5224	2.5282
por	2.2179	2.2141	2.3146
pol	1.9342	1.8868	1.8065
nld	1.7821	1.9149	1.8145
tur	1.2744	1.2328	1.2422
ind	1.1139	1.0532	1.1235
vie	1.0514	1.0332	1.0454
ces	1.0158	1.1019	1.0630

C4 (Colossal Clean Crawled Corpus)

“Очищенный” Common Crawl, представлен в работе про T5

- Оставляли только строки, которые заканчиваются на .?!..
- Предложения минимум из 5 слов, страницы минимум из 3 предложений
- Удалили “нецензурные” тексты (по словарю)
- Удалили все повторы по три предложения
- Оставили только английский!
- Оставили 750-800GB

The Pile

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

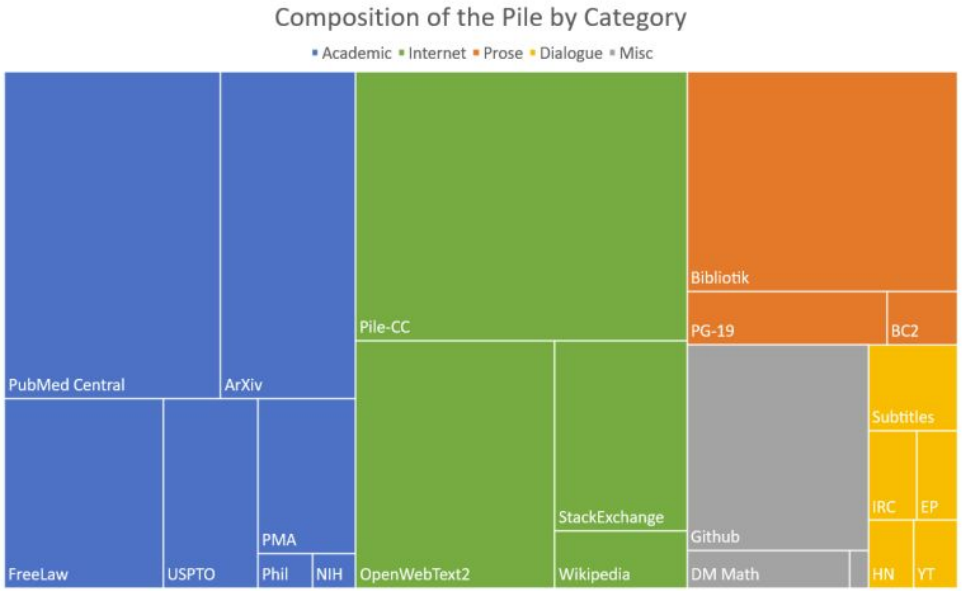


Figure 1: Treemap of Pile components by effective size.

The Pile

- Разнообразный по доменам датасет
- Опять английский язык основной
- ~800GB
- Отдельные методы фильтрации и чистки в зависимости от домена

SlimPajama

- Английский
- 627B токенов
- 900GB сжатым

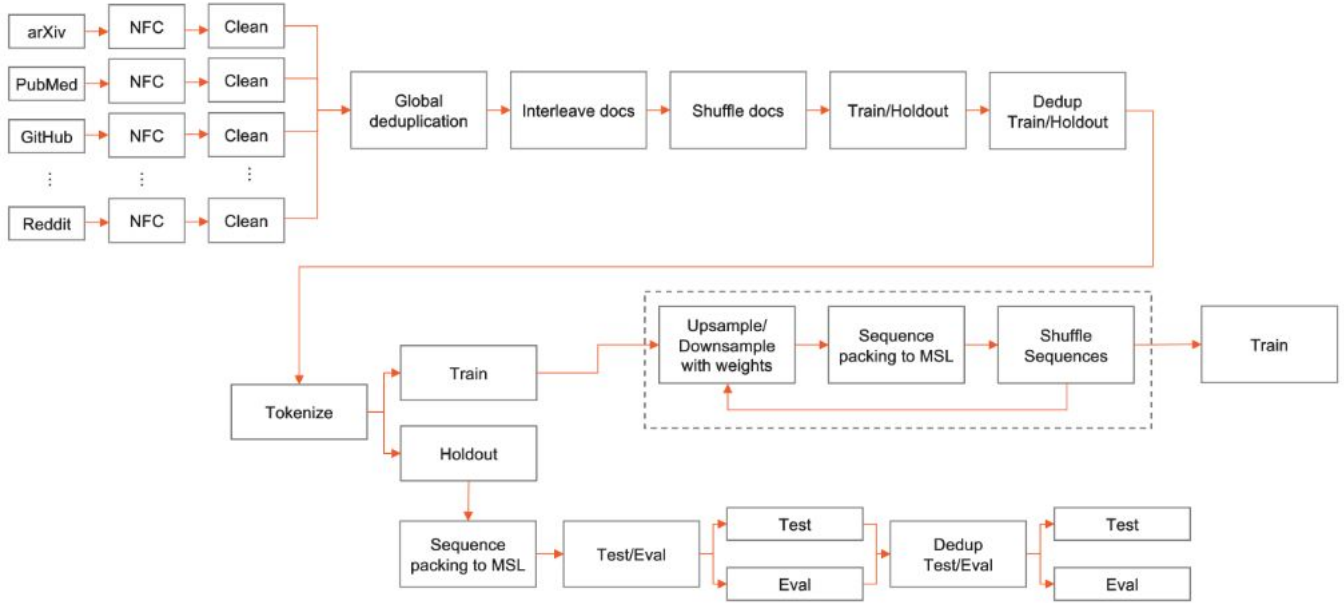
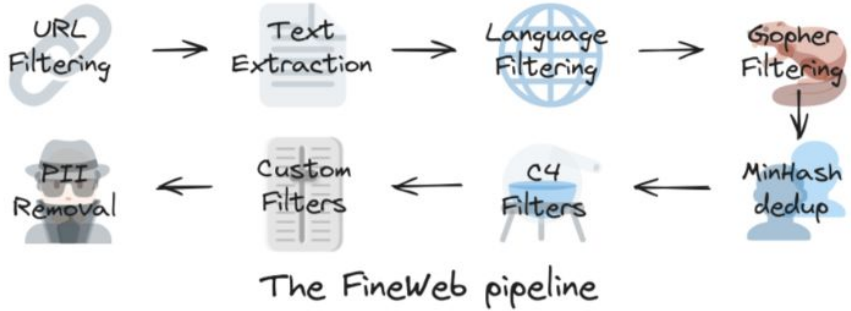


Figure 1: SlimPajama prep-processing pipeline

https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama
https://github.com/Cerebras/modelzoo/tree/e8ab06412e16d3eb82da9a038836a19ce8347b35/src/cerebras/modelzoo/data_preparation/nlp/slimpajama

fineweb

- От Huggingface
- На базе Common Crawl
- 15T токенов, 44 TB на диске
- English
- Есть версия fineweb-edu - качественных образовательных данных (1.3T токенов)



https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1

Omnia Russica

- Новости, Вики, Web
- 90GB
- Чистота / качество данных неизвестны

Omnia Russica

About

Omnia Russica (*lat. all Russian*) is an open source corpus project, containing 33 billion words.

Omnia Russica is combining major Russian corpus sources within one pipeline

	Format	Morphology	Syntax	Size
Wikipedia	vertical	TreeTagger	None	0.5 G
Taiga	CoNLL-U	UDpipe	UDpipe	4.5 G
Araneum Russicum	vertical	TreeTagger	None	25 G
Common Crawl	Plain text	None	None	3 G

rulm

- 75 GB
- Разные домены
- Производилась некоторая чистка

Website	Char count (M)	Word count (M)
pikabu	14938	2161
lenta	1008	135
stihi	2994	393
stackoverflow	1073	228
habr	5112	753
taiga_fontanka	419	55
librusec	10149	1573
buriy	2646	352
ods_tass	1908	255
wiki	3473	469
math	987	177

https://huggingface.co/datasets/IlyaGusev/rulm

Либрусек

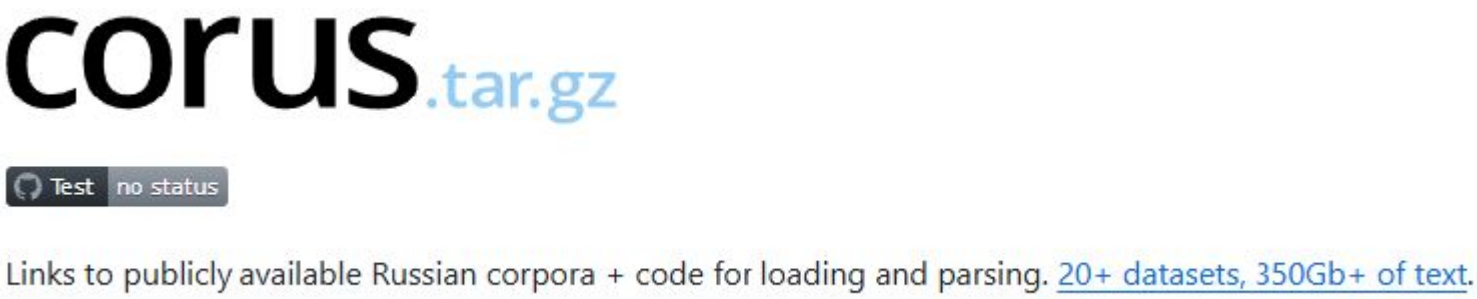
- Лицензия под очень большим вопросом
- Много ГБ данных
- Различные книги, как полезные, так и остальные
- Есть мета-данные

```
"title": datasets.Value("string"),
"file_name": datasets.Value("string"),
"annotation": datasets.Value("string"),
"keywords": datasets.Value("string"),
"date": datasets.Value("string"),
"genre": datasets.Value("string"),
"authors": datasets.Sequence(datasets.Value("string")),
"lang": datasets.Value("string"),
"src_lang": datasets.Value("string"),
"translator": datasets.Value("string"),
"isbn": datasets.Value("string"),
"publisher": datasets.Value("string"),
"city": datasets.Value("string"),
"year": datasets.Value("string"),
"book_name": datasets.Value("string"),
"fancy_title": datasets.Value("string"),
"epigraphs": datasets.Sequence(datasets.Value("string")),
"sections": datasets.Sequence(datasets.Value("string")),
```

https://huggingface.co/datasets/IlyaGusev/librusec_full

Corus

Фреймворк, аккумулирующий разные русский датасеты



https://github.com/natasha/corus

Препроцессинг данных

MassiveWeb пайплайн

Rae J. W. et al. Scaling language models: Methods, analysis & insights from training gopher //arXiv preprint arXiv:2112.11446. – 2021.

MassiveWeb: Content Filtering

Фильтрация очевидно не подходящих документов

- По языку
- По Safety (мат и др.)
- Некоторые ручные правила фильтрации

Фильтрация происходит только на основе информации о документе, можно хорошо параллелить.

MassiveWeb: Text Extraction

- Так как чаще всего на входе Web страницы, нужно извлекать контент.
- То есть нужно парсить HTML разметку, выделять общие паттерны и тп
- Достаточно мало информации про данный этап

MassiveWeb: Quality Filtering

- Большинство Web контента не создано для человека и соответственно для LLM не подходит
- Фильтруют:
  - Документы короче 50 и больше 100 т. слов
  - Средняя длина слова в документе вне диапазона от 3 до 10
  - Отношение количества символов к количеству слов > 0.1
  - Stop word filter
  - …

MassiveWeb: Quality Filtering

Что еще можно добавить

- Фильтрация документов по перплексии
  - Использование маленьких “LLM”
  - Использование классических N-gram LM!
    - Символьные
    - По словам
- Фильтрация документов на основе классификации некоторыми моделями
  - По качеству
  - По приемлемости контента
  - По домену

https://github.com/kpu/kenlm

MassiveWeb: Repetition Removal

Один из показателей данных плохого качества - большое количество повторений.

- Документы с большим количеством повторений внутри удаляются
- Альтернативно можно вырезать из текста слишком явные повторы.

Measurement	Threshold
Duplicate line fraction	0.30
Duplicate paragraph fraction	0.30
Duplicate line character fraction	0.20
Duplicate paragraph character fraction	0.20
Top 2-gram character fraction	0.20
Top 3-gram character fraction	0.18
Top 4-gram character fraction	0.16
Duplicate 5-gram character fraction	0.15
Duplicate 6-gram character fraction	0.14
Duplicate 7-gram character fraction	0.13
Duplicate 8-gram character fraction	0.12
Duplicate 9-gram character fraction	0.11
Duplicate 10-gram character fraction	0.10

MassiveWeb: Document Deduplication

- Крайне важный этап
- Существенный n-gram overlap - убираем
- MinHash algorithm для околодубликатов
  - 13-gramm
  - Jaccard sim
  - Аппроксимация

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Близость двух документов выше 0.8 - удаляем из них
- Игнорируем пунктуацию при расчете n-gram

MassiveWeb: Test-set Filtering

- Мы хотим верить loss на валидации
- Значит надо удалить из Train сета всё слишком близкое к Test-set
- По аналогии с Deduplication

В дополнение

- Некоторые процедуры можно делать in-domain, например, фильтрацию по перплексии
- Для фильтра по языку fasttext lid
- Upsampling и Downsampling
  - Все данные равны, но некоторые, равнее
- Формулы? Таблицы? Latex? …
- Мультимодальные данные?
- Часто огромное количество работы - составление правил и выяснение трешхолдов для них
- В данные нужно смотреть! Garbage in - Garbage out.

Задание

Необходимо создать своей датасет для обучения LLM

- Масштаб от 5GB
- Ориентироваться на качество данных и их чистку
- Потом каждый на своем датасете будет обучать маленькую LLM
- Будет рейтинг итоговых моделей, а качество данных тут играет первоочередную роль
- Результатом выполнения задания являются:
  - Код препроцессинга
  - Сам датасет (его нужно загрузить на HF!)
  - Отчет

Задание: оценивание и сроки

- Срок 1 неделя: до 24 ноября 23:59.
- Присылать на tikhomirov.mm@gmail.com
  - Название письма: LLM & RAG: Задание 5, Данасет
  - В письме Ваше полное ФИО, группа, решение и краткий отчет по нему в PDF.
- Оценка по шкале “-/-+/+-/+/++”.
  - ++ за те решения, которые особо мне понравятся чем-либо.
- Вопросы по заданию можно задавать: в телеграм канале, по почте, в личные сообщения в телеграм.