DataComp

Creating large public Datasets for the AI open-source community

Institute for Foundations of MACHINE LEARNING

Alex Dimakis
UT Austin
Bespoke Labs

Datacomp is a collaboration with Ludwig Schmidt and the DataComp team

Talk overview

- **Datacomp**: A scientific framework for dataset curation.
- 1. Datacomp for multimodal data (Neurips 23)
  2. Datacomp for Language Model Pre-training (DCLM) (Neurips 24)
  3. Ongoing now: Datacomp for Post-Training
- **Philosophical ponderings:** AI systems and the role of synthetic data curation.

Talk overview

Ludwig Schmidt (in this building) told me about LAION. A dataset of billions of images and captions created from Common Crawl.

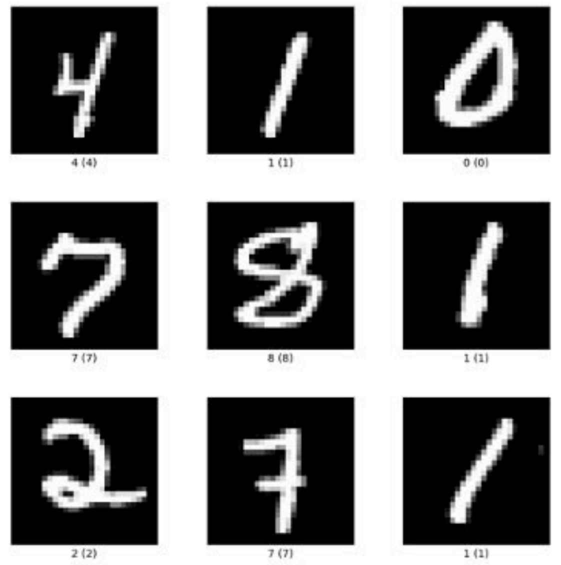
- LAION has been used to train Stable diffusion, and many other multimodal models. Also used for most open CLIP model pipelines.
- Idea: Make a better version of the LAION dataset. But also:
  Open-source the **tooling** for dataset curation.
  Create **scientific standardized benchmarks** for dataset curation
  Create a **community**.

Datasets are the foundation of progress in AI

- **For text:**

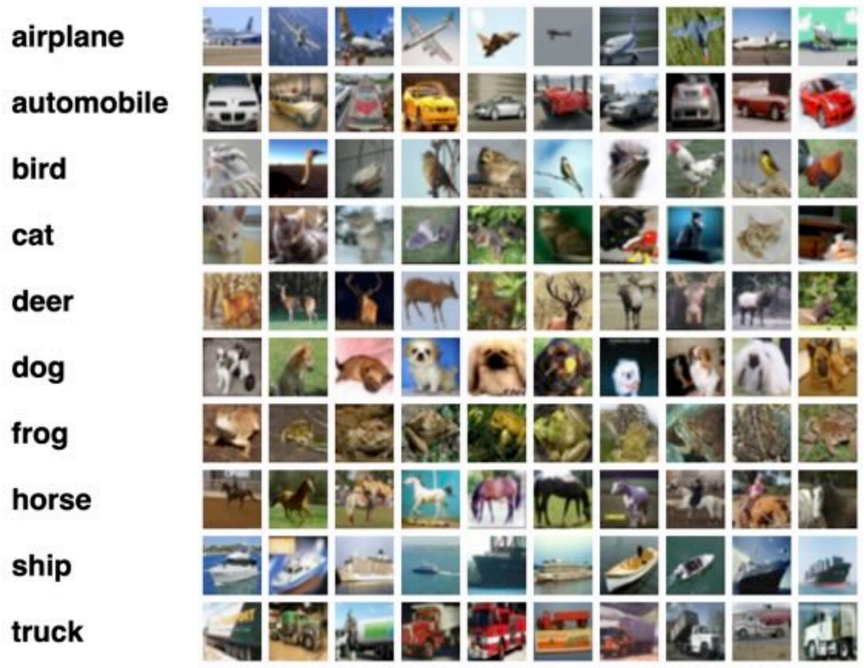
- GPT-1 (2018): 3 B Tokens
- GPT-2 (2019): 30 B Tokens
- GPT-3 (2020): 300 B Tokens
- GPT-4 (2013): 3000? B Tokens
- **1000x growth in 5 years**
- **For images:**
- Imagenet (2009): 1 Million images
- LAION-5B (2022): 5 Billion Images
- **5000x growth in 5 years**

ML Discoveries enabled by datasets



MNIST (1994)

Convolutional neural networks



ImageNet (2012)

Deep learning resurgence, ResNets, transfer learning, etc.

OpenAI

WebImageText (2021)

Zero-shot classification

Data is a poorly understood (yet crucial) component of LLMs

LLMs rely on **trillions of tokens** of training data crawled from the Internet.

Details about training sets have become sparse for state of the art models.

Amongst open-source models, significant gap between closed v.s. open dataset models.

Average of 22 standard downstream evaluations

Model	Params	Tokens	Open dataset?	CORE	MMLU
Open weights, closed datasets					
Llama2	7B	2T	✗	49.2	45.8
DeepSeek	7B	2T	✗	50.7	48.5
Mistral-0.3	7B	?	✗	57.0	62.7
QWEN-2	7B	?	✗	57.5	**71.9**
Llama3	8B	15T	✗	57.6	66.2
Gemma	8B	6T	✗	57.8	64.3
Phi-3	7B	?	✗	**61.0**	69.9
Open weights, open datasets					
Falcon	7B	1T	✓	44.1	27.4
OLMo-1.7	7B	2.1T	✓	47.0	54.0
MAP-Neo	7B	4.5T	✓	**50.2**	**57.1**

Data is a poorly understood (yet crucial) component of LLMs

LLMs rely on **trillions of tokens** of training data crawled from the Internet.

Details about training sets have become sparse for state of the art models.

Amongst open-source models, significant gap between closed v.s. open dataset models.

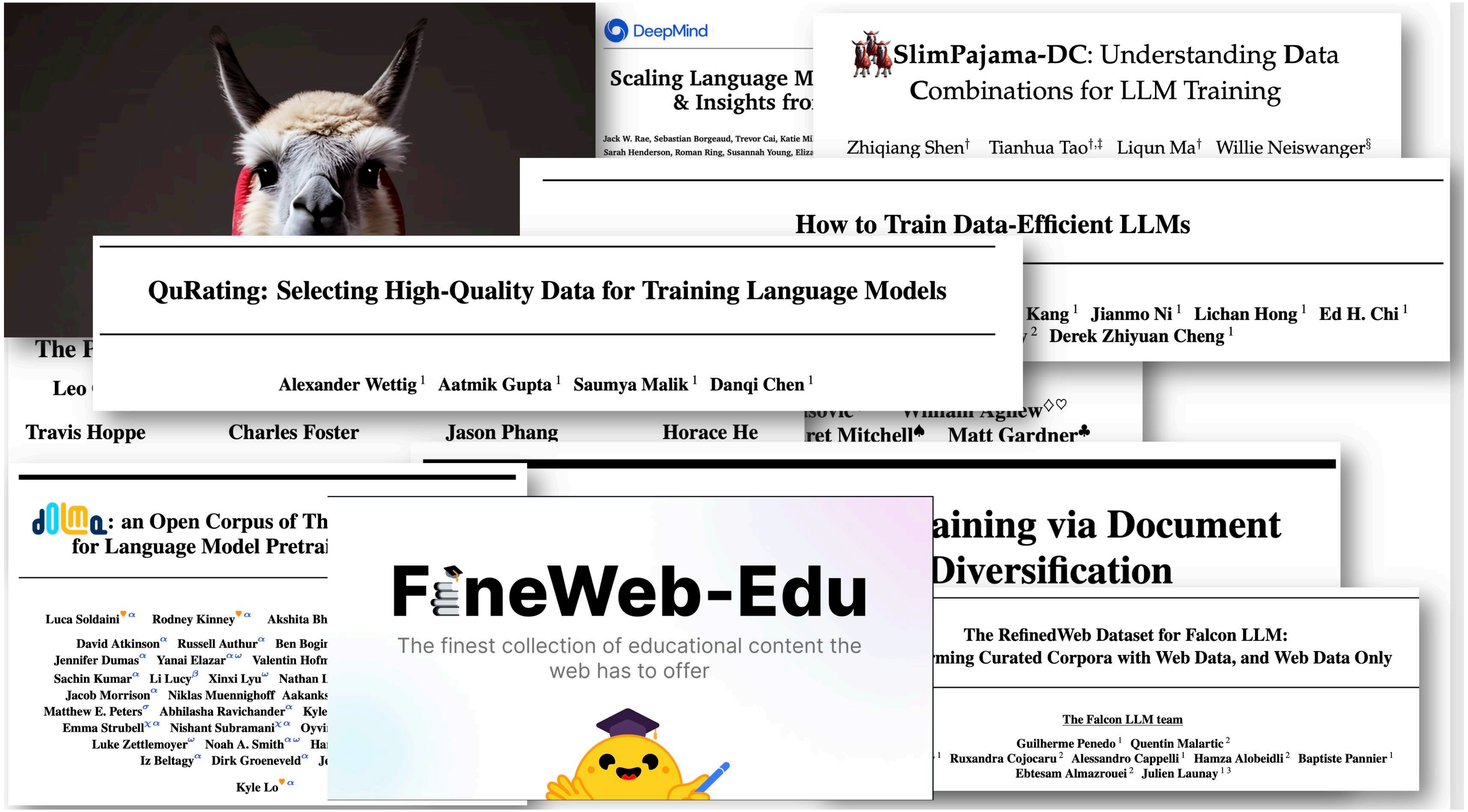
Average of 22 standard downstream evaluations

Model	Params	Tokens	Open dataset?	CORE	MMLU
Open weights, closed datasets					
Llama2	7B	2T	✗	49.2	45.8
DeepSeek	7B	2T	✗	50.7	48.5
Mistral-0.3	7B	?	✗	57.0	62.7
QWEN-2	7B	?	✗	57.5	**71.9**
Llama3	8B	15T	✗	57.6	66.2
Gemma	8B	6T	✗	57.8	64.3
Phi-3	7B	?	✗	**61.0**	69.9
Open weights, open datasets					
Falcon	7B	1T	✓	44.1	27.4
OLMo-1.7	7B	2.1T	✓	47.0	54.0
MAP-Neo	7B	4.5T	✓	**50.2**	**57.1**

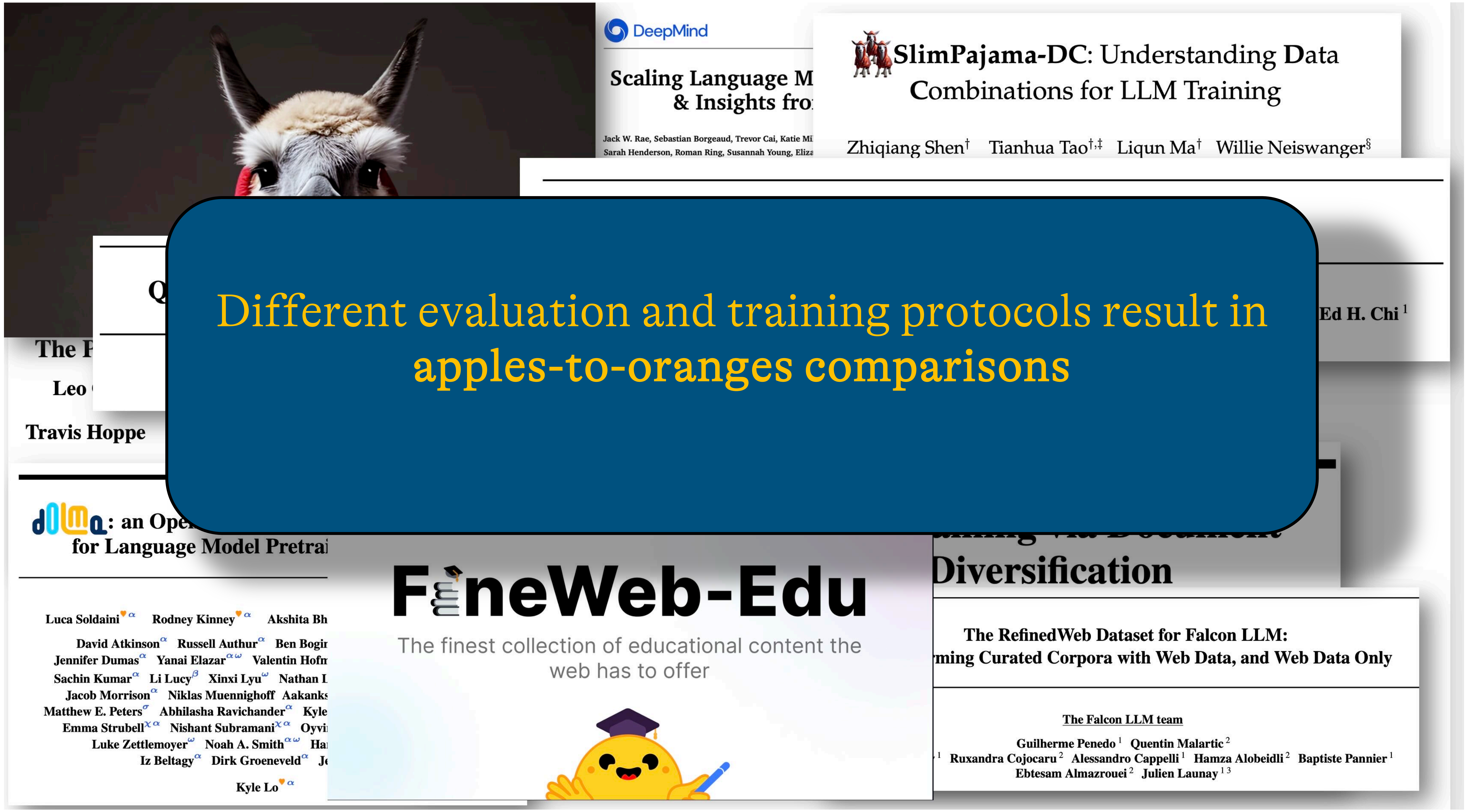
Spoiler:

We got an open data model with MMLU 64

A problem with previous work on dataset curation



A problem with previous work on dataset curation



Different evaluation and training protocols result in apples-to-oranges comparisons

The DataComp idea: Shift to Data-Centric AI

Traditional ML Benchmark: Dataset fixed (e.g. Imagenet),

improve the model (Model Centric)

DataComp: Model training +Evaluation fixed.

Improve the dataset. (Data Centric)

DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2] Gabriel Ilharco*[1] Alex Fang*[1] Jonathan Hayase[1] Georgios Smyrnis[5]
Thao Nguyen[1] Ryan Marten[7,10] Mitchell Wortsman[1] Dhruba Ghosh[1] Jieyu Zhang[1] Eyal Orgad[3]
Rahim Entezari[11] Giannis Daras[5] Sarah Pratt[1] Vivek Ramanujan[1] Yonatan Bitton[12]
Kalyani Marathe[1] Stephen Mussmann[1] Richard Vencu[6] Mehdi Cherti[6,8,9] Ranjay Krishna[1]
Pang Wei Koh[1] Olga Saukh[11] Alexander Ratner[1] Shuran Song[2] Hannaneh Hajishirzi[1,7]
Ali Farhadi[1] Romain Beaumont[6] Sewoong Oh[1] Alexandros G. Dimakis[5] Jenia Jitsev[6,8,9]
Yair Carmon[3] Vaishaal Shankar[4] Ludwig Schmidt[1,6,7]

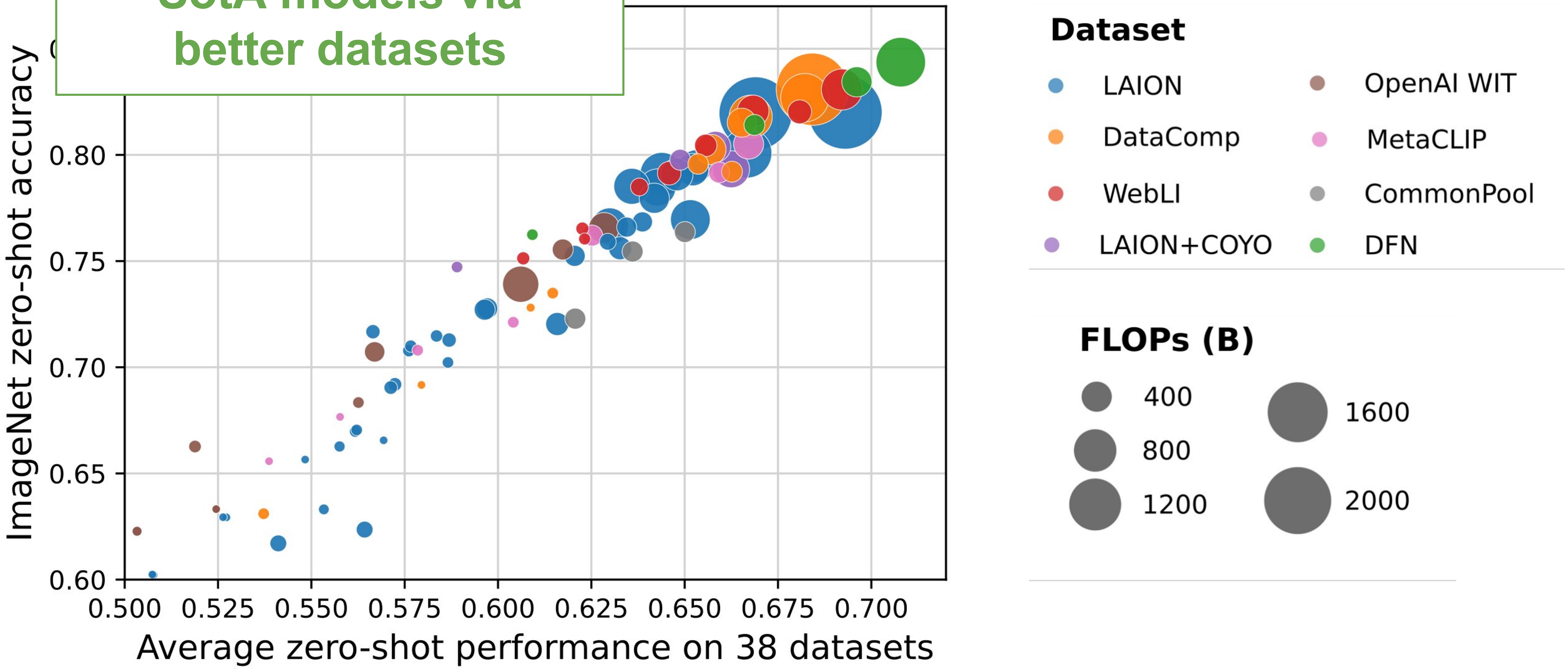
Abstract

Large multimodal datasets have been instrumental in multiple breakthroughs like CLIP, DALL-E, Stable Diffusion, Flamingo and GPT-4, yet datasets rarely receive the same research attention as model architectures or training algorithms. To address this shortcoming in the machine learning ecosystem, we introduce DataComp, a participatory benchmark where the training code is fixed and researchers innovate by proposing new training sets. Concretely, we provide an experimental testbed centered around a new

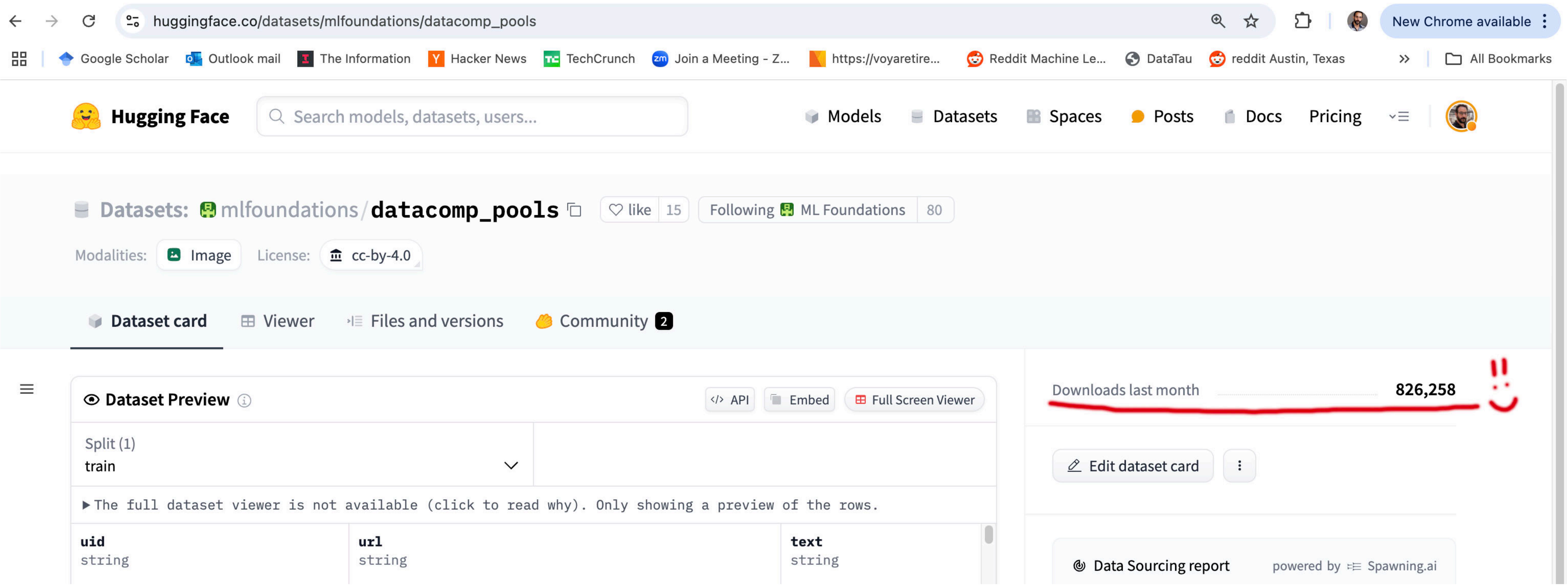
University of Washington
Tel Aviv University
Columbia University
Jülich Forschungszentrum
Texas
The University of Texas at Austin
Ai2
האוניברסיטה העברית בירושלים
The Hebrew University of Jerusalem
IFML
University of Illinois Urbana-Champaign
TU Graz
stability.ai
Google

Impact of DataComp

SotA models via better datasets



Impact of DataComp



DataComp - LM

In search of the next generation of training sets for language models

The DCLM Team

- **Alex Fang** University of Washington
- **Jeffrey Li** University of Washington
- **Georgios Smyrnis** UT Austin
- **Maor Ivgi** Tel Aviv University
- **Matt Jordan** UT Austin
- **Samir Gadre** Columbia University
- **Gabriel Ilharco** University of Washington
- **Giannis Daras** UT Austin
- **Kalyani Marathe** University of Washington
- **Aaron Gokaslan** Cornell University
- **Jieyu Zhang** University of Washington
- **Khyathi Chandu** AI2

- **Hritik Bansal** UCLA
- **Etash Guha** University of Washington
- **Sedrick Keh** Toyota Research Institute
- **Kushal Arora** Toyota Research Institute
- **Saurabh Garg** Carnegie Mellon University
- **Rui Xin** University of Washington
- **Thao Nguyen** University of Washington
- **Igor Vasiljevic** Toyota Research Institute
- **Sham Kakade** Harvard University
- **Shuran Song** Stanford University
- **Sujay Sanghavi** UT Austin
- **Fartash Faghri** Apple

- **Niklas Muennighoff** Contextual AI
- **Reinhard Heckel** Technical University of Munich
- **Jean Mercat** Toyota Research Institute
- **Mayee Chen** Stanford University
- **Suchin Gururangan** University of Washington
- **Mitchell Wortsman** University of Washington
- **Sewoong Oh** University of Washington
- **Luke Zettlemoyer** University of Washington
- **Kyle Lo** AI2
- **Alaa El-Nouby** Apple
- **Hadi Pour Ansari** Apple
- **Alexander Toshev** Apple

- **Alon Albalak** UCSB
- **Yonatan Bitton** Hebrew University
- **Marianna Nezhurina** Research Center Juelich & JSC & LAION
- **Amro Abbas** Datology AI
- **Cheng-Yu Hsieh** University of Washington
- **Dhruba Ghosh** University of Washington
- **Stephanie Wang** University of Washington
- **Dirk Groeneveld** AI2
- **Luca Soldaini** AI2
- **Pang Wei Koh** University of Washington
- **Jenia Jitsev** Research Center Juelich & JSC & LAION
- **Thomas Kollar** Toyota Research Institute

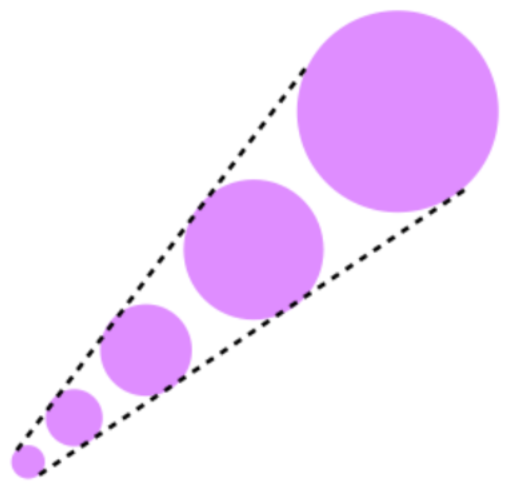
- **Joshua Gardner** University of Washington
- **Maciej Kilian** USC
- **Hanlin Zhang** Harvard University
- **Rulin Shao** University of Washington
- **Sarah Pratt** University of Washington
- **Sunny Sanyal** UT Austin
- **Alex Dimakis** UT Austin BespokeLabs.ai
- **Yair Carmon** Tel Aviv University
- **Achal Dave** Toyota Research Institute
- **Ludwig Schmidt** University of Washington, Stanford
- **Vaishaal Shankar** Apple

DCLM Contributions

- A **benchmark** for language model data curation
- **416 experiments** to identify the key parts of an effective data curation pipeline
- A **state-of-the-art public training set** for 7B parameter models
- A **leaderboard** for the **community** to try different data curation methods
- Our best model, **DCLM 7B** achieves 64 MMLU, trained only on 2T tokens.
- This is better than Llama2 models while Llama3 8B was trained on 6x times more tokens (i.e. 6 times more compute efficient, due to better data curation!).

DCLM benchmark

A. Select a scale



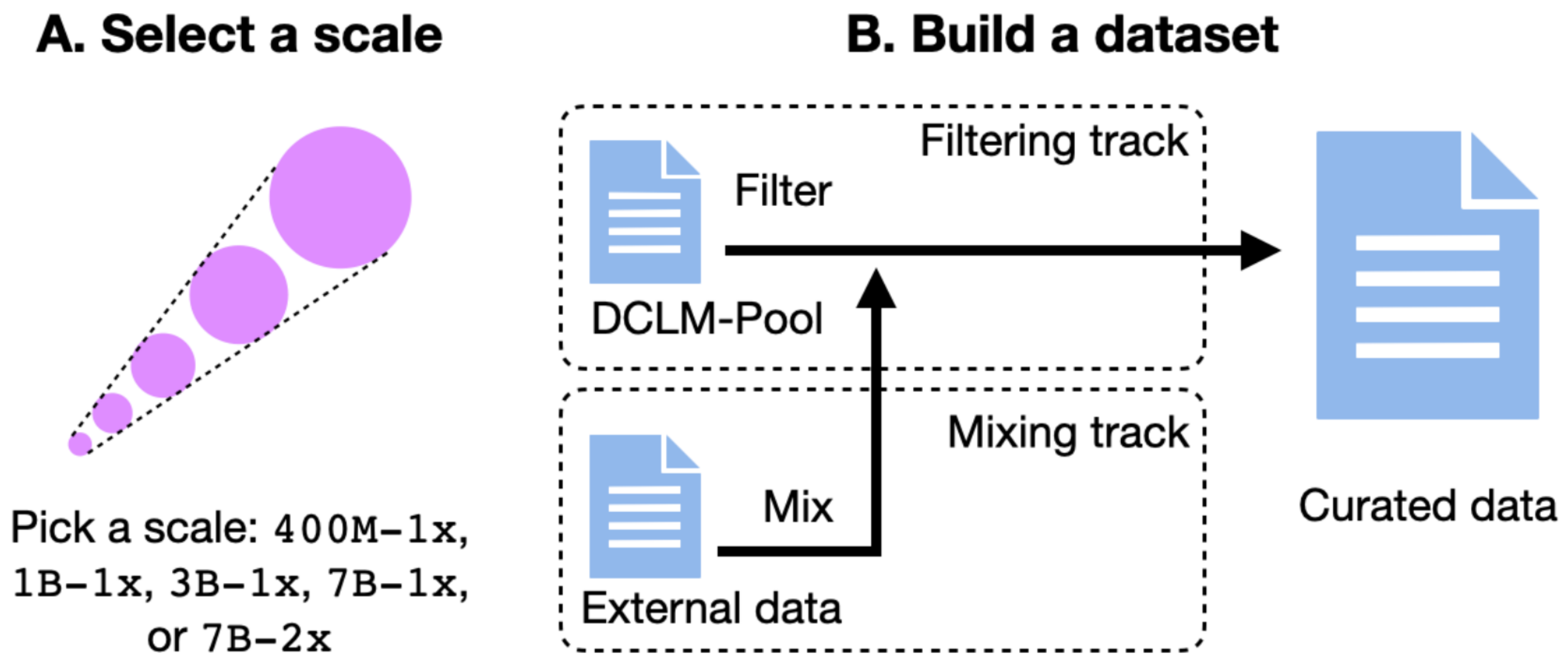
Pick a scale: `400M-1x`, `1B-1x`, `3B-1x`, `7B-1x`, or `7B-2x`

Scale	Model parameters	Train tokens	Train FLOPs	Train H100 hours	Pool size
400M-1x	412M	8.2B	2.0e19	26	469B
1B-1x	1.4B	28.8B	2.4e20	240	1.64T
3B-1x	2.8B	55.9B	9.4e20	740	3.18T
7B-1x	6.9B	138B	5.7e21	3,700	7.85T
7B-2x	6.9B	276B	1.1e22	7,300	15.7T

We fix initial data pools, training recipes, and evaluations.

Participants submit datasets that can then be fairly compared to one another.

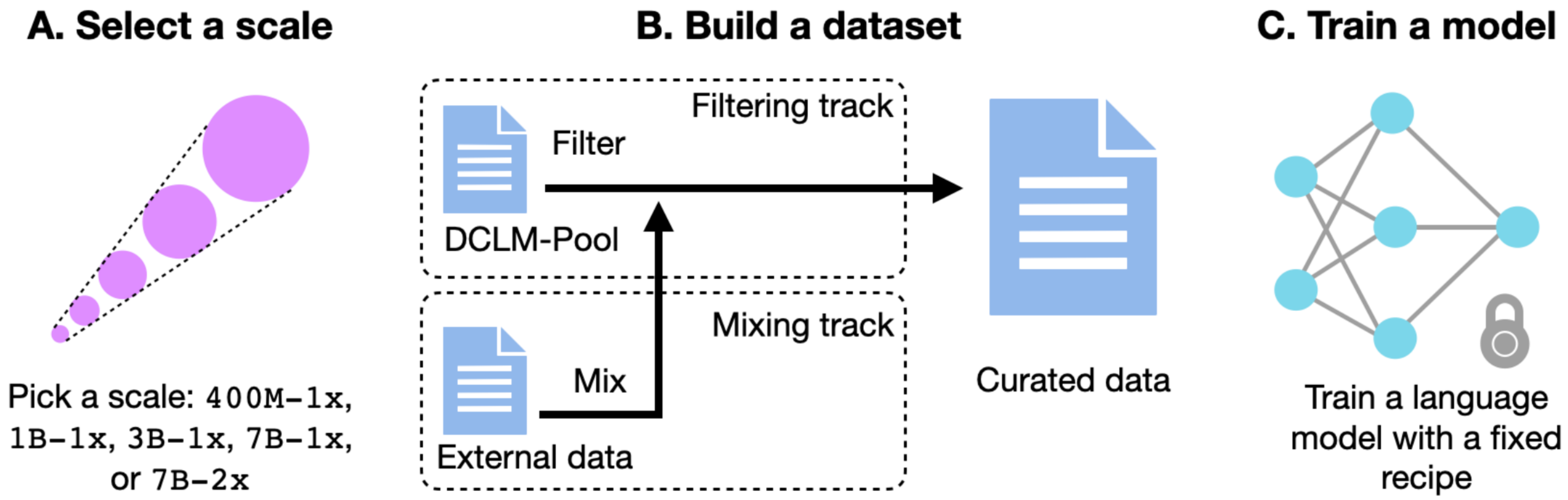
DCLM benchmark



We fix initial data pools, training recipes, and evaluations.

Participants submit datasets that can then be fairly compared to one another.

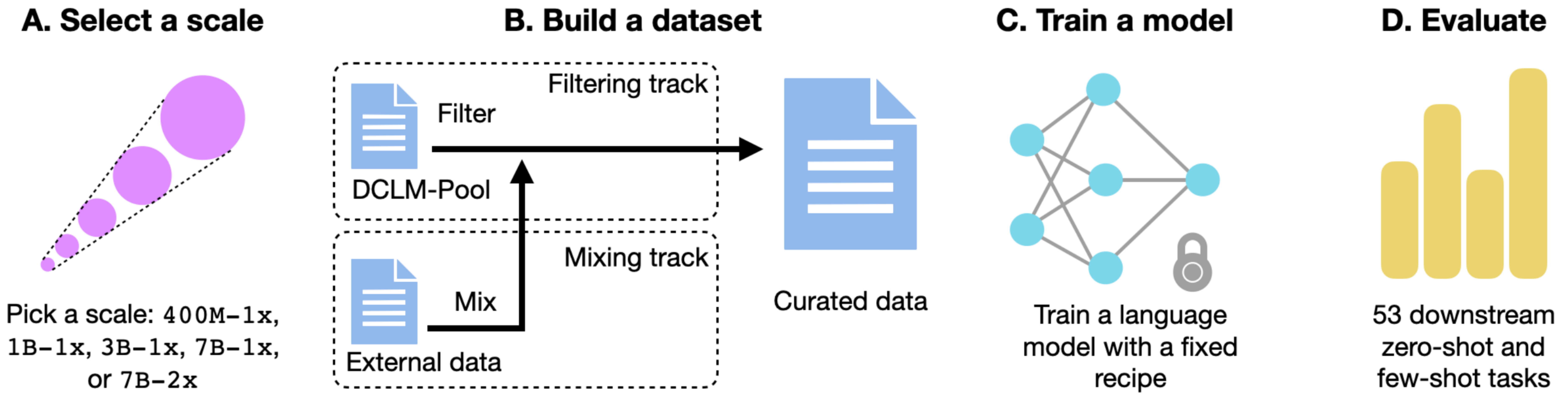
DCLM benchmark



We fix initial data pools, training recipes, and evaluations.

Participants submit datasets that can then be fairly compared to one another.

DCLM benchmark



We fix initial data pools, training recipes, and evaluations.

Participants submit datasets that can then be fairly compared to one another.

Lessons from DCLM

1. Data curation algorithms that work at 400M scale
   can indeed predict performance at bigger sizes (at least up to 7B).

2. That means we can do data curation science by developing data curation algorithms at small scale and extrapolating.

3. You don't have to build the whole ship and throw it in the ocean to check if it floats. Build a minature ship and test in the bathtub.

Lessons from DCLM

1. Deduplication and data cleaning are very important.
(Bloom filters, Min-hash, a lot more to do here)

2. Humans are terrible at deciding which paragraphs will be good vs bad pretraining data. We found this very surprising.

3. The best data filter we found is a FastText Classifier trained on Instruction tuning vs CommonCrawl.
This outperformed everything else we tried!

4. Once you have created a good-quality filtered dataset, no mixing methods seemed to help further which was surprising.
(Right ways to Ensemble datasets?)

Effective data filtering

We use a combination of the **RefinedWeb heuristic filters** + a **trained classifier**.

[Penedo et al., 2023]

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [176]	27.2	14.0
AskLLM [139]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
`fastText` [81] OH-2.5 +ELI5	**30.2**	**15.4**

`Results for 1B-1x scale (1.4B model, ~28B tokens)`

→ A `fastText` bigram classifier **combined with the right training data** does best

Training data is key for the data filtering model

Dataset	Threshold	CORE	MMLU
OH-2.5 + ELI5	10%	**41.0**	**29.2**
Wikipedia	10%	35.7	27.0
OpenWebText2	10%	34.7	25.0
GPT-3 Approx	10%	37.5	24.4
OH-2.5 + ELI5	15%	39.8	27.2
OH-2.5 + ELI5	20%	38.7	24.2

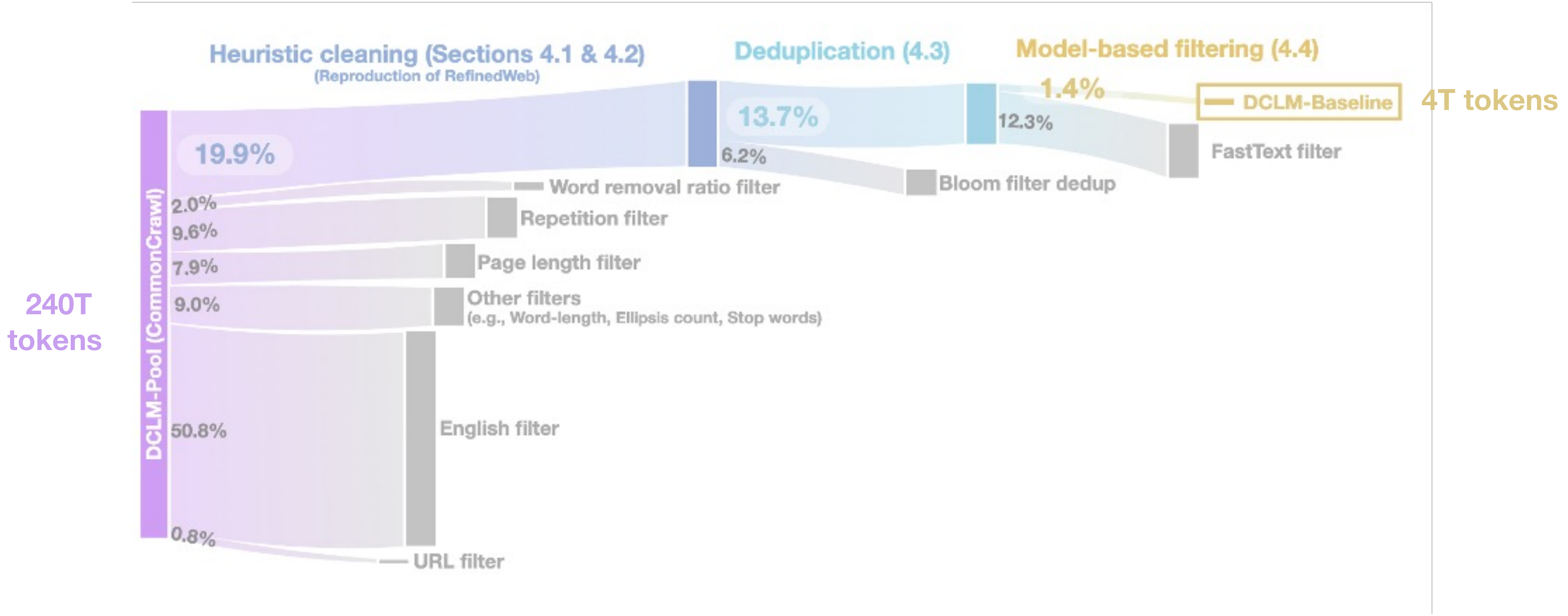
Results for 7B-1x scale (7B model, ~140B tokens)

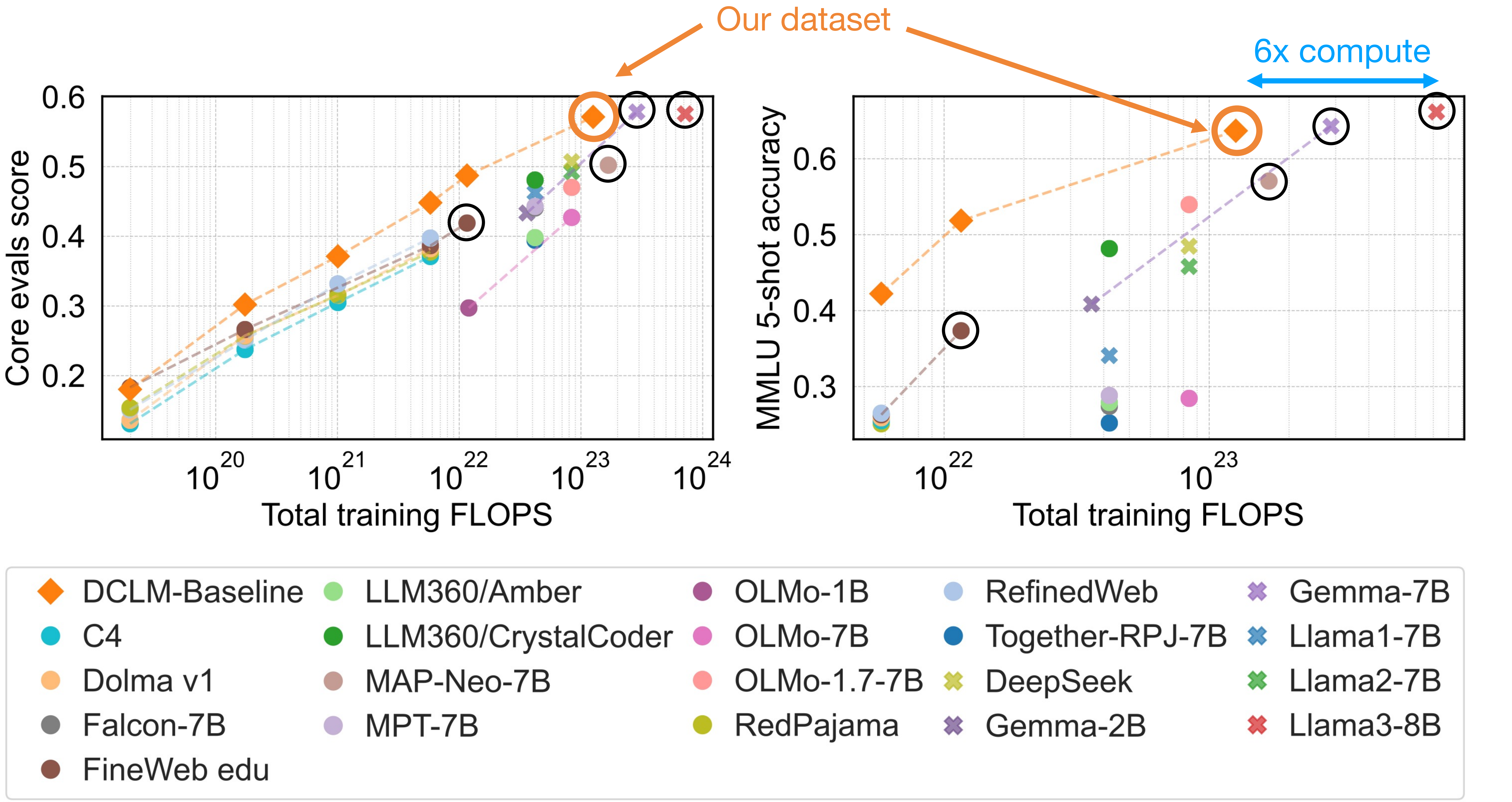
- **OpenHermes-2.5:** a SoTA instruction tuning dataset (mix of ~15 sources)
- **ELI5:** pairs of questions and top answers from the /r/explainlikeimfive subreddit

>

Previous sources (GPT-3 Approx):
Wiki + Books + OpenWebText2

Putting it all together: DCLM-Baseline





DCLM-Baseline outperforms leading open-source datasets like FineWeb-Edu and MAP-Neo and closes gap to **Mistral-7B**, Gemma-7B, and Llama 3-8B (with 6x less compute).

Check out the DCLM paper for more findings!


DataComp-LM: In search of the next generation of training sets for language models

Jeffrey Li*[1,2] Alex Fang*[1,2] Georgios Smyrnis*[4] Maor Ivgi*[5] Matt Jordan[4] Samir Gadre[3,6] Hritik Bansal[8] Etash Guha[1,15] Sedrick Keh[3] Kushal Arora[3] Saurabh Garg[13] Rui Xin[1] Niklas Muennighoff[22] Reinhard Heckel[12] Jean Mercat[3] Mayee Chen[7] Suchin Gururangan[1] Mitchell Wortsman[1] Alon Albalak[19,20] Yonatan Bitton[14] Marianna Nezhurina[9,10] Amro Abbas[23] Cheng-Yu Hsieh[1] Dhruba Ghosh[1] Josh Gardner[1] Maciej Kilian[17] Hanlin Zhang[18] Rulin Shao[1] Sarah Pratt[1] Sunny Sanyal[4] Gabriel Ilharco[1] Giannis Daras[4] Kalyani Marathe[1] Aaron Gokaslan[16] Jieyu Zhang[1] Khyathi Chandu[11] Thao Nguyen[1] Igor Vasiljevic[3] Sham Kakade[18] Shuran Song[6,7] Sujay Sanghavi[4] Fartash Faghri[2] Sewoong Oh[1] Luke Zettlemoyer[1] Kyle Lo[11] Alaaeldin El-Nouby[2] Hadi Pouransari[2] Alexander Toshev[2] Stephanie Wang[1] Dirk Groeneveld[11] Luca Soldaini[11] Pang Wei Koh[1] Jenia Jitsev[9,10] Thomas Kollar[3] Alexandros G. Dimakis[4,21] Yair Carmon[5] Achal Dave†[3] Ludwig Schmidt†[1,7] Vaishaal Shankar†[2]

[1]University of Washington, [2]Apple, [3]Toyota Research Institute, [4]UT Austin, [5]Tel Aviv University, [6]Columbia University, [7]Stanford, [8]UCLA, [9]JSC, [10]LAION, [11]AI2, [12]TUM, [13]CMU, [14]Hebrew University, [15]SambaNova, [16]Cornell, [17]USC, [18]Harvard, [19]UCSB, [20]SynthLabs, [21]Bespokelabs.AI, [22]Contextual AI, [23]DatologyAI

contact@datacomp.ai

Abstract

We introduce DataComp for Language Models (DCLM), a testbed for controlled dataset experiments with the goal of improving language models. As part of DCLM, we provide a standardized corpus of 240T tokens extracted from Common Crawl, effective pretraining recipes based on the OpenLM framework, and a broad suite of 53 downstream evaluations. Participants in the DCLM benchmark can experiment with data curation strategies such as deduplication, filtering, and data mixing at model scales ranging from 412M to 7B parameters. As a baseline for DCLM, we conduct extensive experiments and find that model-based filtering is key to assembling a high-quality training set. The resulting dataset, DCLM-BASELINE, enables training a 7B parameter language model from scratch to 64% 5-shot accuracy on MMLU with 2.6T training tokens. Compared to MAP-Neo, the previous state-of-the-art in open-data language models, DCLM-BASELINE represents a 6.6 percentage point improvement on MMLU while being trained with 40% less compute. Our baseline model is also comparable to Mistral-7B-v0.3 and Llama 3 8B on MMLU (63% & 66%), and performs similarly on an average of 53 natural language understanding tasks while being trained with 6.6× less compute than Llama 3 8B. Our results highlight the importance of dataset design for training language models and offer a starting point for further research on data curation. We release the DCLM benchmark, framework, models, and datasets at https://datacomp.ai/dclm.

- Resiliparse and Bloom Filters are effective for extraction and deduplication
- Dataset rankings are consistent across scales and hyperparameters
- Adding traditional “high-quality” sources (e.g., Wikipedia) doesn’t always help
- Contamination is unlikely to explain the strong performance of DCLM-Baseline
- Agreement with human labels isn’t a good predictor among high-quality filters

Let's all improve open-source datasets!

DataComp   Home   FAQs   Team   Leaderboard

DataComp - LM

Welcome to DataComp, the machine learning benchmark where the models are fixed and the challenge is to find the best possible data!

Prior competitions in machine learning have focused on finding the best model, with a fixed set of training and test data. However, many recent advances (GPT-4, Gemini, LLAMA, Mistral) are due in part to large and diverse language datasets. DataComp centers the role that data plays by fixing the training code, and encouraging researchers to innovate by proposing new training sets. We provide an experimental testbed centered around a new candidate pool of 240T tokens from Common Crawl. Participants in our benchmark design new filtering techniques or curate new data sources and then evaluate them by running our standardized language model training code followed by an evaluation on 53 downstream datasets. Our benchmark consists of multiple scales, with various candidate pool sizes and associated compute budgets ranging from 412M to 7B parameters. This multi-scale design facilitates the study of scaling trends and makes the benchmark accessible to researchers with varying resources.

Paper   Code   Data

Rank	Created	Submission	Core	Extended	MMLU	Authors	Writeup
1	06-19-2024	Baseline: DCLM-Baseline	0.44823	0.28776	0.42228	DCLM team	https://arxiv.org/abs/24
2*	06-19-2024	External: RefinedWeb	0.39749	0.21665	0.24765	The Falcon LLM team	https://arxiv.org/abs/23
3*	06-19-2024	External: Fineweb-Edu	0.38658	0.22062	0.26294	HuggingFace	https://huggingface.co
4*	06-19-2024	External: Dolma v1	0.38220	0.19573	0.25890	AI2	https://arxiv.org/abs/24
5*	06-19-2024	External: RPJ	0.37752	0.20614	0.25111	Together AI	https://www.together.a
6*	06-19-2024	External: C4	0.37103	0.19840	0.25505	Google	https://arxiv.org/abs/19

dclm Public   Edit Pins   Watch 38   Fork 104   Starred 1.2k

COMMON CRAWL

Dataset Name	Size	Number of Objects	Description
DCLM-pool	279.6 TiB	5,047,684	Raw extracted text
DCLM-refinedweb	41.6 TiB	357,163	Raw text subsequently processed with a pipeline similar to RefinedWeb
DCLM-baseline	6.6 TiB	27,938	RefinedWeb filtered using ML

https://www.datacomp.ai/dclm/

How this project incorporates foundational work

The idea of data centric AI is a new paradigm.

We need to design new AI models for filtering. We need theoretical principles for data quality filtering, data subset selection etc.

Evals are hard. Just released Evalchemy, an LLM evaluation platform today.

A lot of on-going interest on training with synthetic data. Several new foundational questions arise and Datacomp can be a benchmark for synthetic datasets.

On-going: Datacomp for post-training: Create a summarization dataset by prompting GPT. Post-train a small (8B Llama) models on synthetic instruction datasets.

Theoretical principles for foundation model specialization, using GPT4o as an Oracle.

Small Specialized Models through Synthetic Data curation

Paradigm 1: The GPT 5 AGI monolith.
A gigantic model that knows everything.
You prompt it to summarize papers.
Students prompt it to ask questions about your lecture notes etc.

Small Specialized Models through Synthetic Data curation

Paradigm 1: The GPT 5 AGI monolith.
A gigantic model that knows everything.
You prompt it to summarize papers.
Students prompt it to ask questions about your lecture notes etc.

Paradigm 2: Small specialized models
Can be created by synthetic dataset curation:
1.You prompt the big model to create an instruction dataset:
1. Summarize these 10k Neurips papers.
2.Check if this synthetic dataset is truthful, faithful, helpful etc.
3.Distill a small LLM to make a Neurips paper summarizer
4. Evaluate that this is working well

The small model can be better than its teacher (e.g. as shown in Bespoke-Minicheck 7B)
(if the prompting pipeline uses external knowledge, external validators or competition).

- Fin