A Data Centric LLMs

Hyemi Kim

March 27, 2025

B9145: Topics in Trustworthy AI

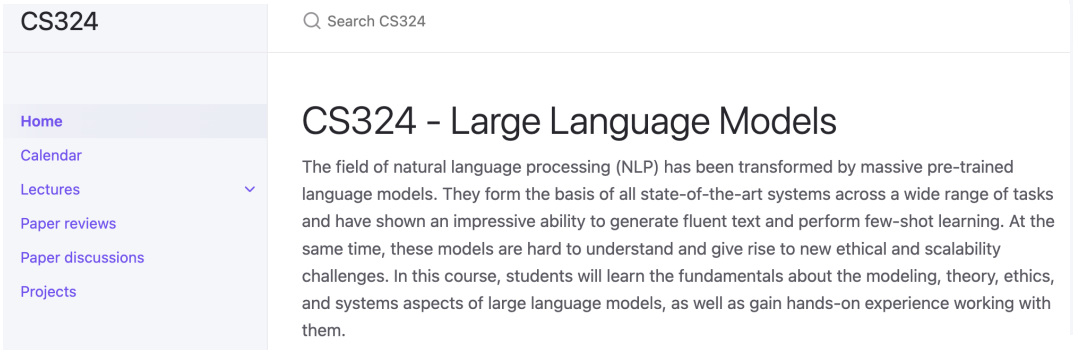
Just Let You Know…

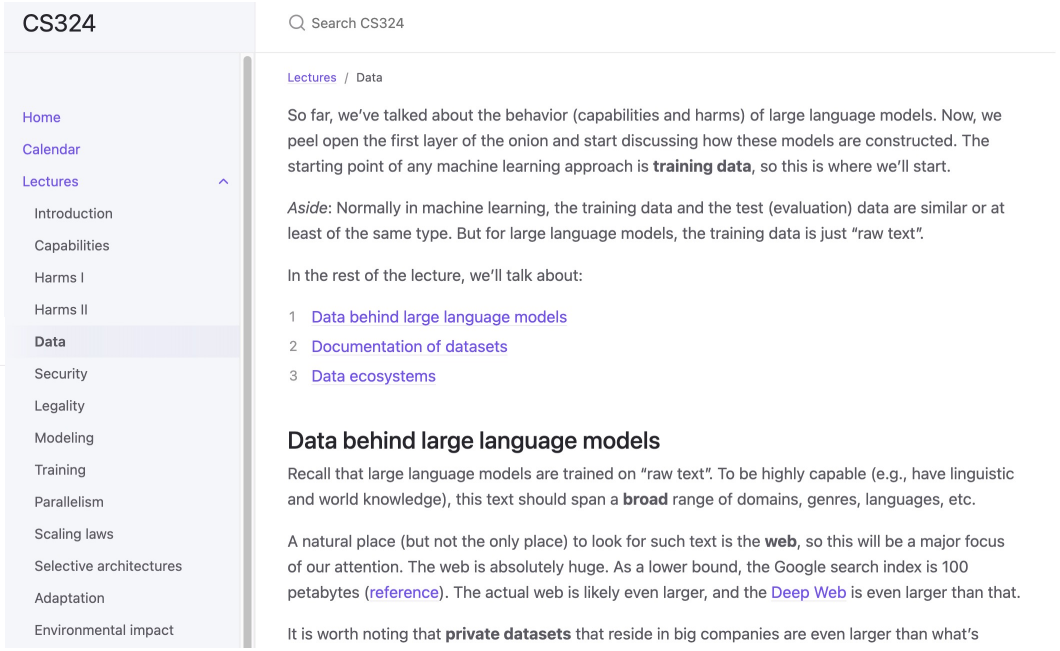
- Less (almost no) *theory*, more ***anecdotal stories*** about *the societal impacts of LLMs*, or even ***guidelines***…
- **Goal**: After this presentation, I'd be happy if even just one of the following happens,
  - You recognize the importance of taking a holistic view of the data ecosystems
  - You can recall at least one piece of evidence showing that LLMs might be toxic or unethical
  - If you have learned the key guidelines for documentation
  - … or at least one slide sticks with you (except this slide)

Contents

A. Data Ecosystems
B. Data Behind LLMs
C. Data Curation

Heavily rely on…

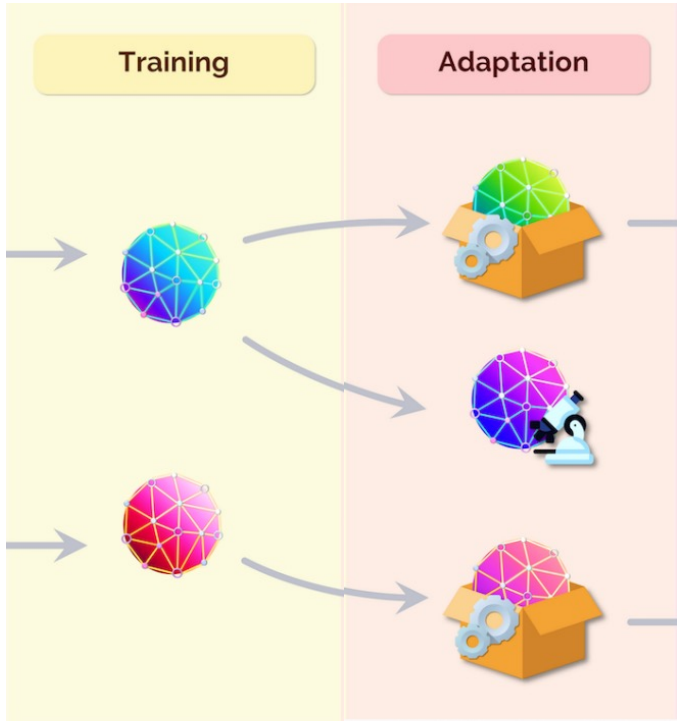




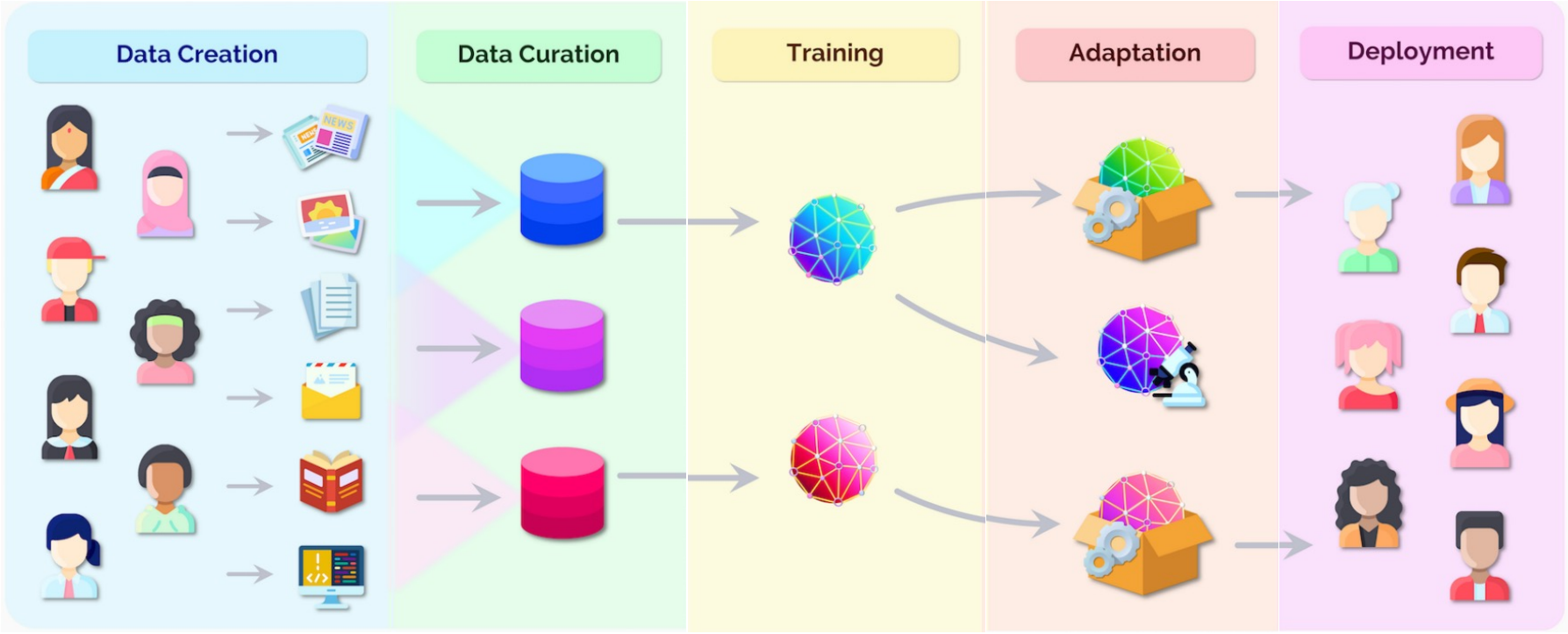
+ some recent papers

A. Data Ecosystems

Where do LLMs inhabit?

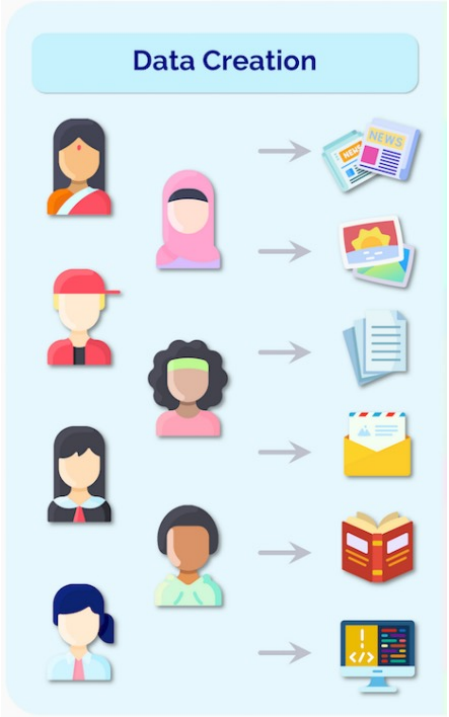


Data Ecosystems



Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

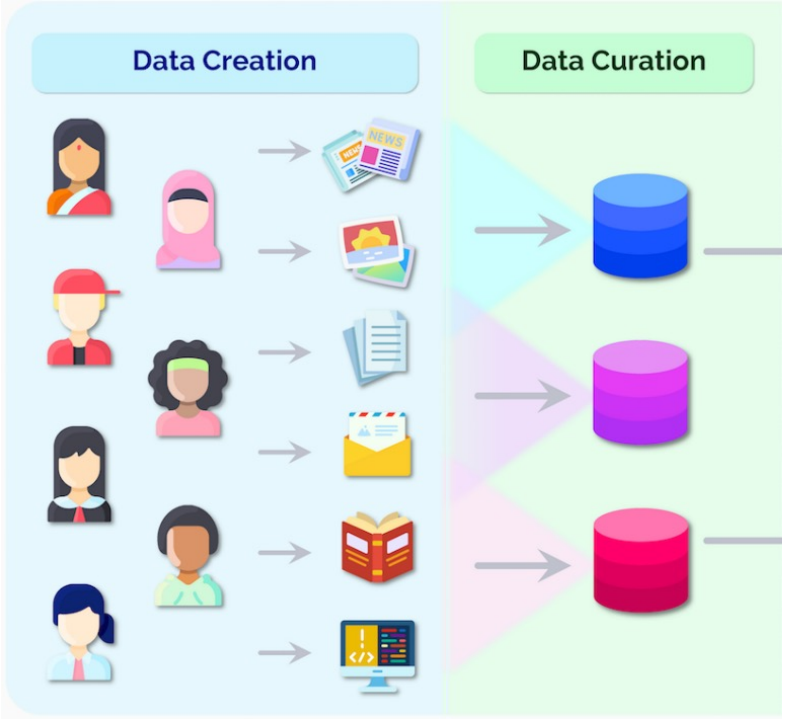
(1) Data Creation



- All data is created by **people** and about **people**
  - Data can be a **measurement of** people (e.g., genomic data)
  - Data can be a **measurement of the environment** (e.g., satellite images)
- All data has **an owner** and **a purpose**

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

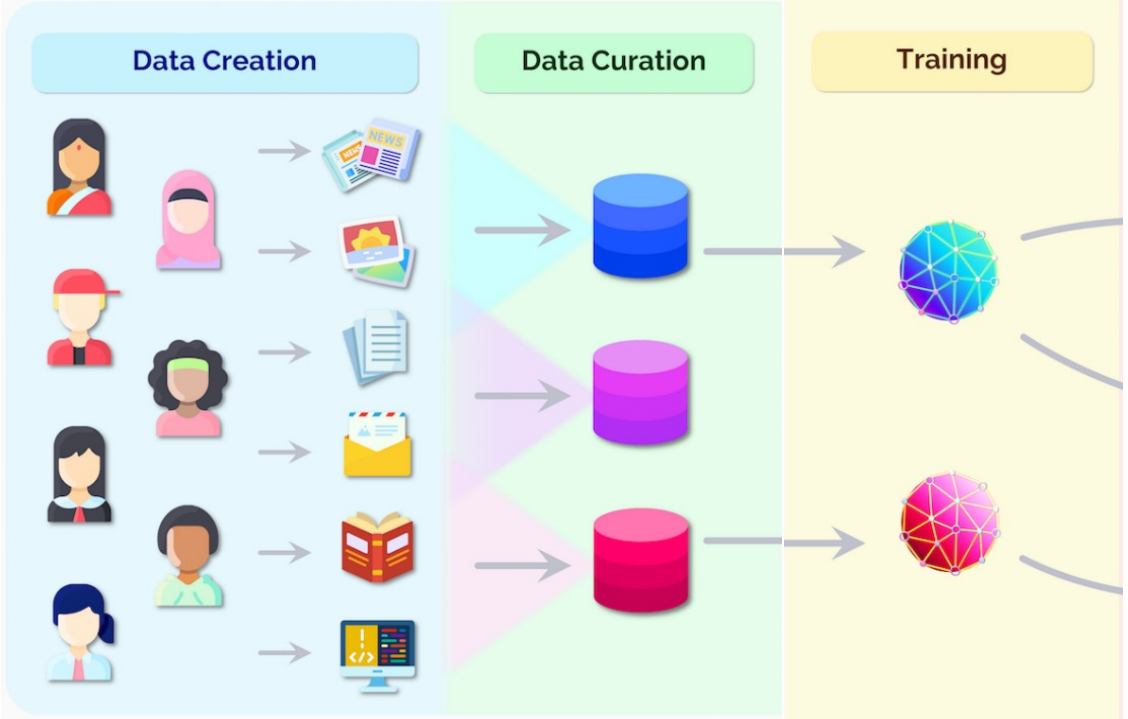
(2) Data Curation



- Every dataset involves **selection** and **filtering**
- Ensuring **relevance** and **quality** while respecting **legal** and **ethical** constraints
- Industry prioritizes it, but AI research often overlooks it

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

(3) Training

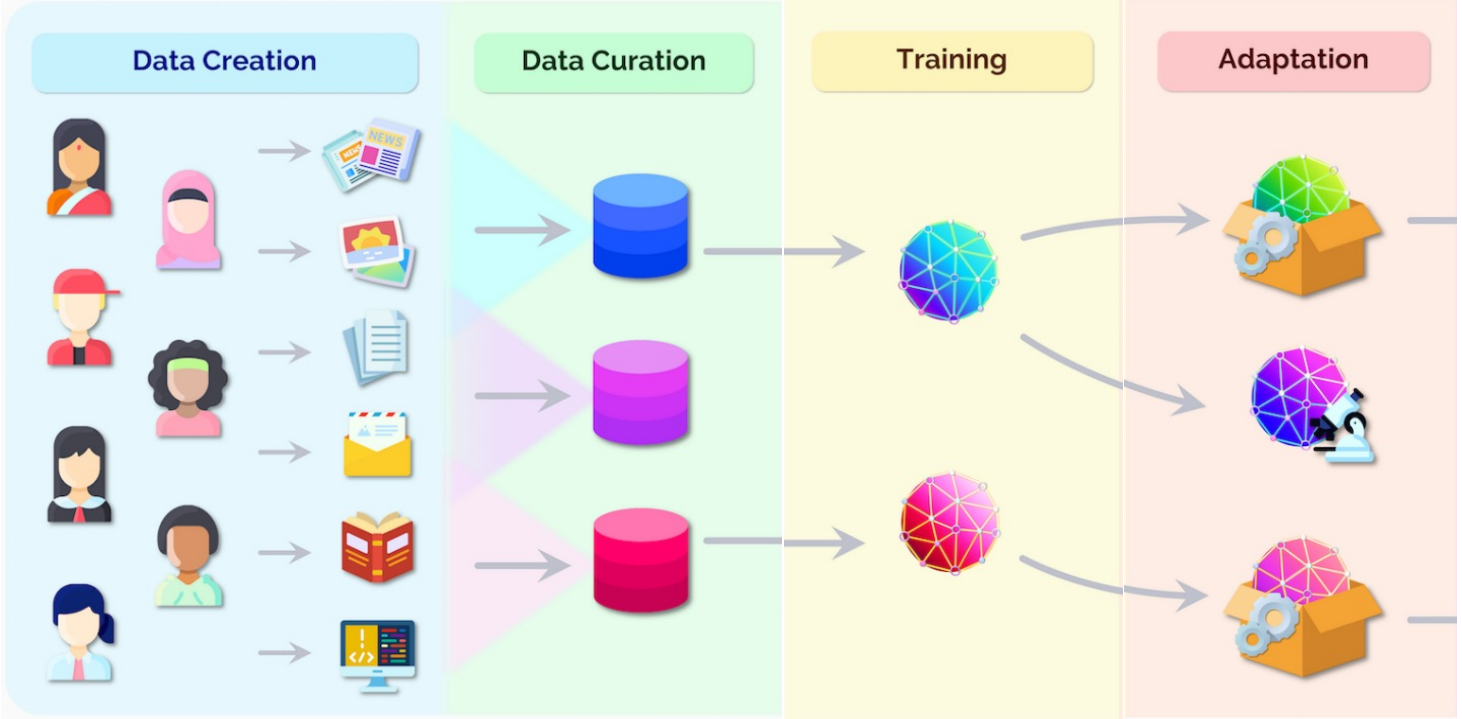


- Training LLMs relies on curated datasets
- It is a key focus in AI research but ***only one stage of many!***

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

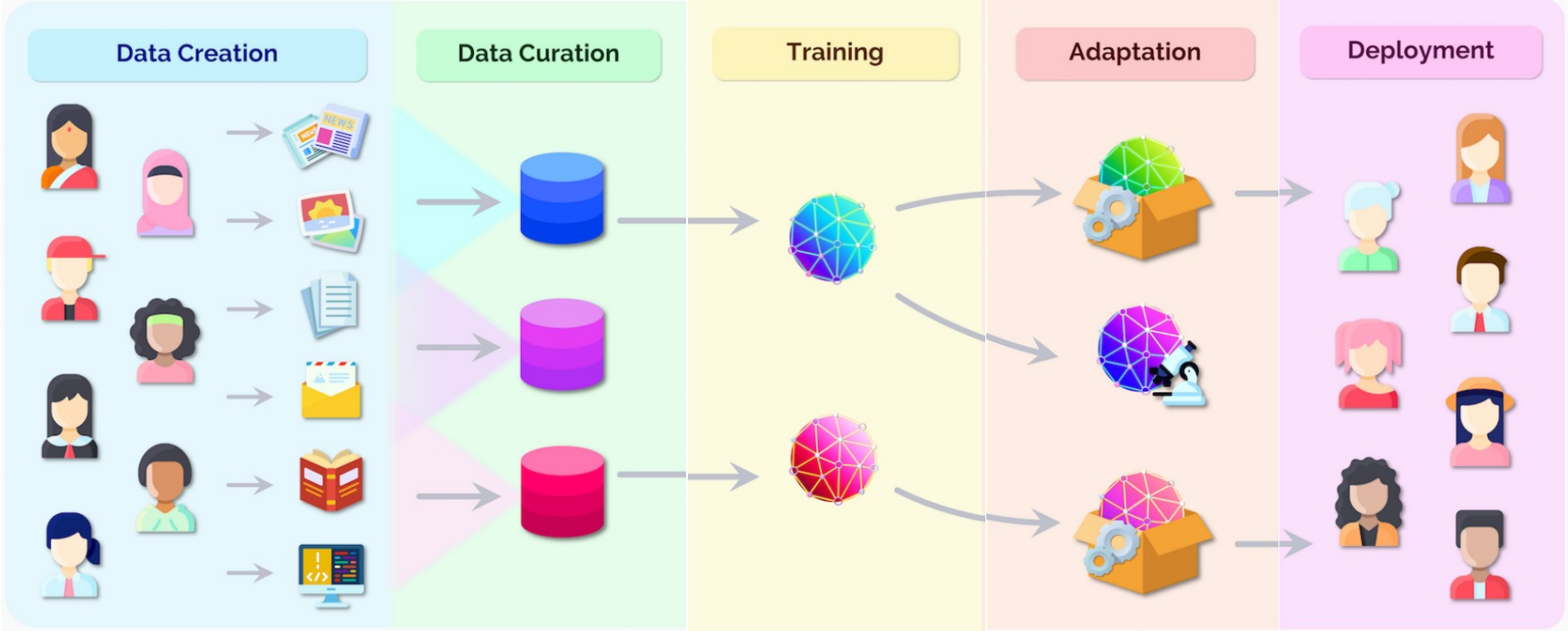
Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

(4) Adaptation



- Some applications require customization before use
- **How?** Two main methods:
  - Adding new data or task-specific prompts (e.g., *TL;DR* for summarization)
  - Updating model parameters (**fine-tuning**) with domain-specific data
- **E.g.,** Task specialization, domain adaptation, test-time data removal, etc.

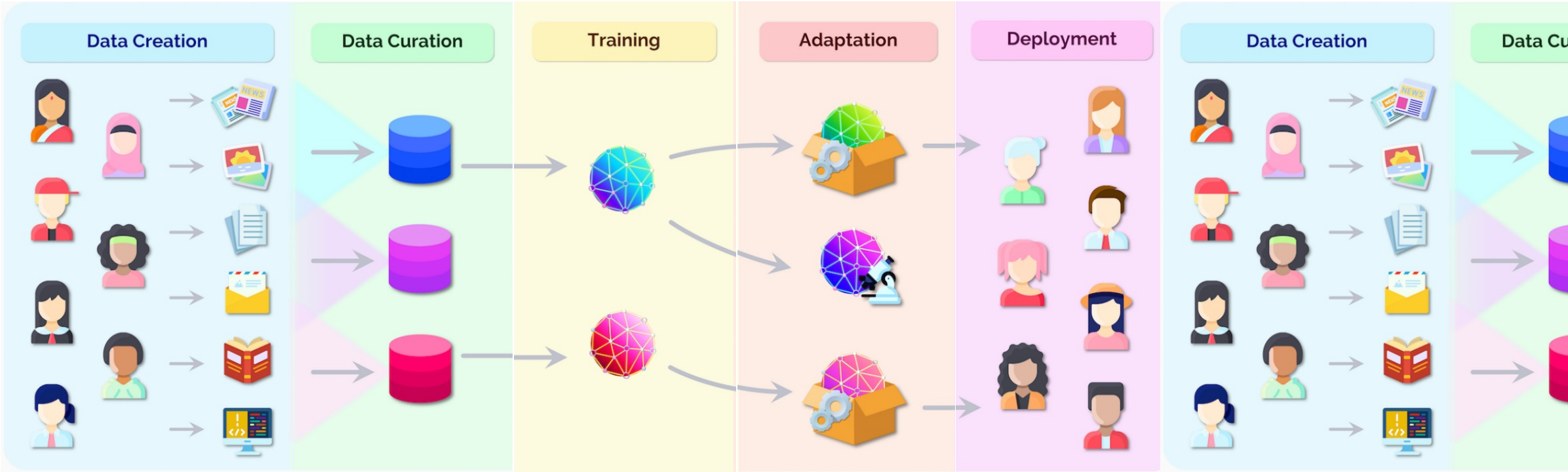
(5) Deployment

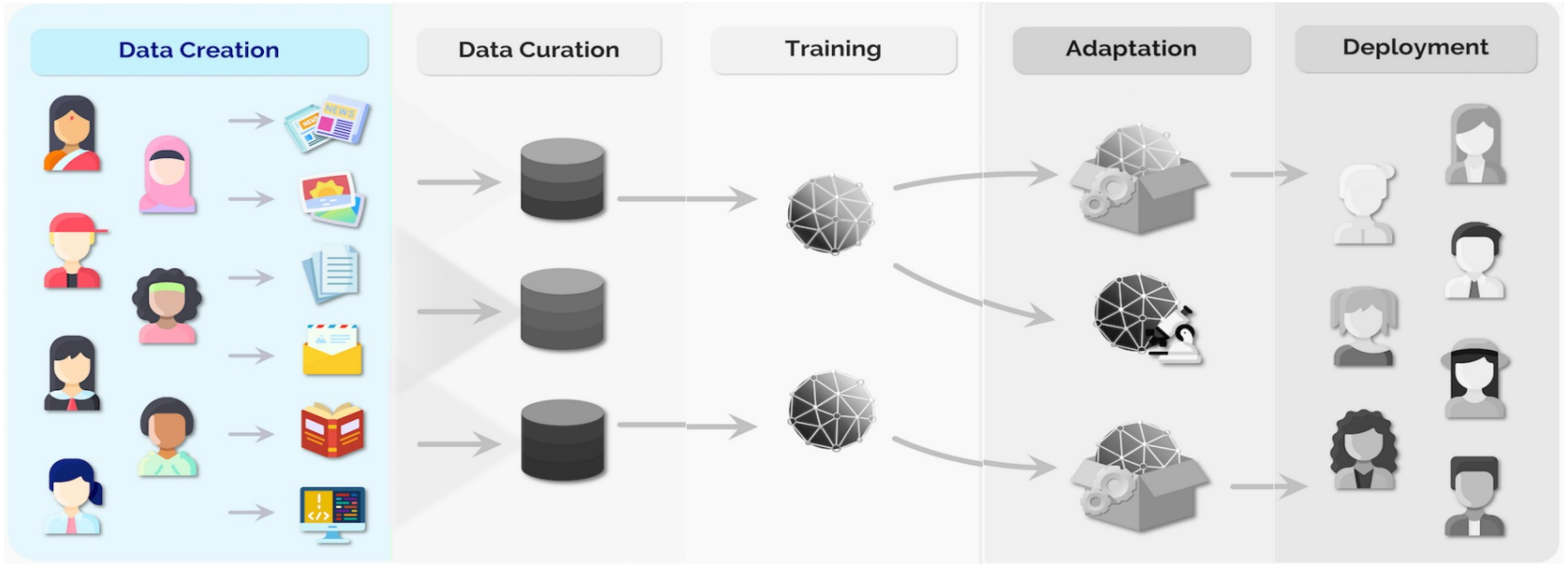


- Direct effects occur when real users interact with the system
- Harmful models shouldn’t be deployed but but may have research value
- Gradual testing helps catch issues before wide release

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

(+) Feedback…?


Data Creation
Data Curation
Training
Adaptation
Deployment
Data Creation
Data Cu



B. Data Behind LLMs

& Its Societal Impacts

Massive Data

- LLM models are trained on “raw text”
  - The text should span a broad range of domains, genres, languages, etc.
- One of the places to look for such text is **web**
  - The actual web $\geq$ The Google Search Index $\geq$ 100 PB
  - 100 PB $\approx$ 1M 4K-movies $\approx$ 10 trillion copies of the Harry Potter series
- Private datasets: Walmart generates 2.5PB/hour

Massive Data ≠ Diversity

On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

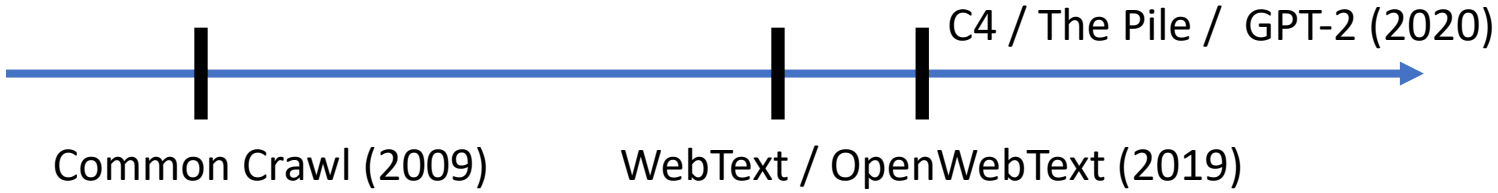
Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

It has been noted in Bender et al, 2021 that despite the size, large-scale data still has **uneven representation** over the population:

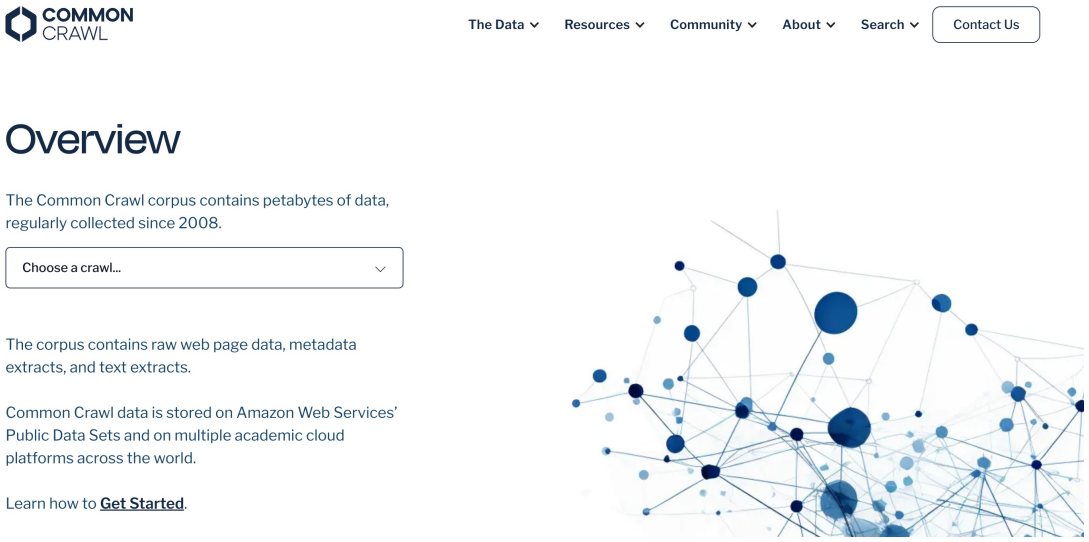
- Internet data overrepresents younger users from developed countries
- GPT-2's training data is based on *Reddit*, which according to Pew Internet Research's 2016 survey, 67% of Reddit users are men, 64% between ages 18 and 29 in the US
- Only 8.8-15% of Wikipedians are female
- Harassment on Internet could turn away certain people (trans, queer, neurodivergent people)
- Filtering "bad words" could further marginalize certain populations (e.g., LGBTQIA+)

Some datasets (or not) for LLMs:



Common Crawl – Introduction (1/4)

- **Common Crawl** is a nonprofit organization that crawls the web and provides snapshots that are free to the public
- A standard source of data to train models such as T5, GPT-3, and Gopher
- 320 TB data (April 21)



- WARC Files (Web ARChive Format)
  - Contains full raw HTML and HTTP responses of web pages
  - Includes images, JavaScript, and other embedded content
- WAT Files (Web Archive Transform)
  - Metadata extracted from WARC files
  - Includes page structure, headers, and links
- WET Files (Web Extracted Text)
  - Contains only the extracted text from HTML pages (without HTML tags)

Common Crawl – Problems (2/4)

- Luccioni et al. (2021) find that Common Crawl contains a significant amount of undesirable content, including hate speech and sexually explicit content, even after filtering procedures
- "… Unfortunately, the majority of the resulting text is not natural language. Instead, it largely comprises *gibberish* or *boiler-plate text like menus, error messages, or duplicate text*. Furthermore, a good deal of the scraped text contains content that is unlikely to be helpful for any of the tasks we consider (offensive language, placeholder text, source code, etc.)…" [Raffel, Colin, et al. (2020)]

Luccioni, Alexandra Sasha, and Joseph D. Viviano. "What's in the box? a preliminary analysis of undesirable content in the common crawl corpus." *arXiv preprint arXiv:2105.02732* (2021)
Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.

Understanding Common Crawl (3/4)

- Baack (2024) conducts a qualitative analysis on Common Crawl
- "Common Crawl its data *wants* to contain *problematic content* to enable open-ended research and innovation”

> “You know, why do you need Common Crawl? It’s all out there on the web, you can just go get it yourself. But it’s difficult to start and operate a web crawler, so if you’re a researcher and you want to do some kind of study but need a billion pages before you can start, that’s a lot of work and there are a lot of issues involved with that.” (Interview CC director)

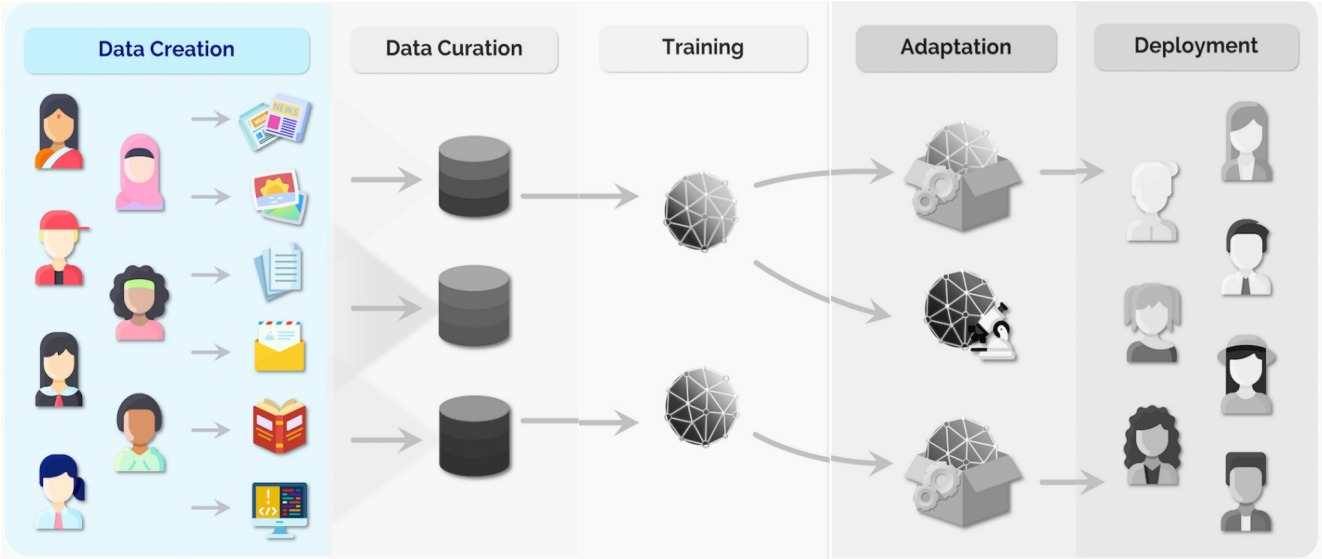
The absence of content curation or moderation is framed as vital to this infrastructural quality. As the director put it, less curation enables more research and open innovation by downstream users:

> “From a goal standpoint, I don’t think we want to necessarily be curating the dataset because the pages we removed might be of value to downstream users. You might be looking for the prevalence of hate speech within a certain country. . .if you’re the researcher trying to measure the prevalence, you want that material in there. So we kind of said it’s sort of up to the downstream user to do content classification.” (Interview CC director)

Baack, Stefan. "A critical analysis of the largest source for generative ai training data: Common crawl." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024.

Understanding Common Crawl (3/4)

- Baack (2024) conducts a qualitative analysis on Common Crawl
- "Common Crawl its data *wants* to contain *problematic content* to enable open-ended research and innovation, but it does not want to take responsibility for annotating it”



Baack, Stefan. "A critical analysis of the largest source for generative ai training data: Common crawl." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024.

To Whom Will Use Common Crawl (4/4)

1. Avoid Misconception

Please do not think "Common Crawl is the copy of internet"

> "That's something I try to explain to everyone: Often it is claimed that Common Crawl contains the entire web, but that's absolutely not true. Based on what I know about how many URLs exist, it's very, very small. I think that's really important." (Interview CC crawl engineer)

- Big and important domains like *New York Times/Facebook/etc.* block the *Common Crawl*
- Majority of content in Common Crawl is English (all the technical infrastructure is based in the US!)
- The director describes Common Crawl as an "academic sampling of the web" (use harmonic centrality)

To Whom Will Use Common Crawl (4/4)

1. Avoid Misconception

Please do not think “Common Crawl is the copy of internet”

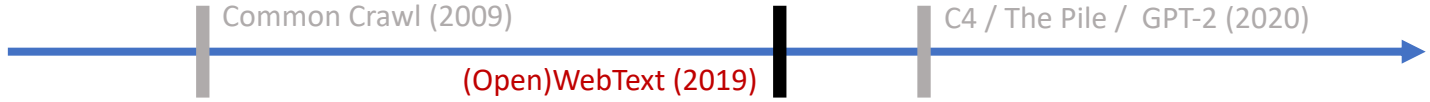
2. Stronger Content Filtering

Beyond removing pornographic content

3. Diversify/Tailor Data Source

Do not over-rely on specific Common Crawl versions (e.g. C4, Pile-CC)

4. …



WebText and OpenWebText (1/4)

WebText: used to train GPT-2 (not released by OpenAI) [Redford et al. (2019)]

- Scraped all outbound links from *Reddit* that received at least 3 karma (upvotes)
- Filtered out *Wikipedia* to be able to evaluate on Wikipedia-based benchmarks
- End result is 40 GB of text after de-duplication and some heuristics

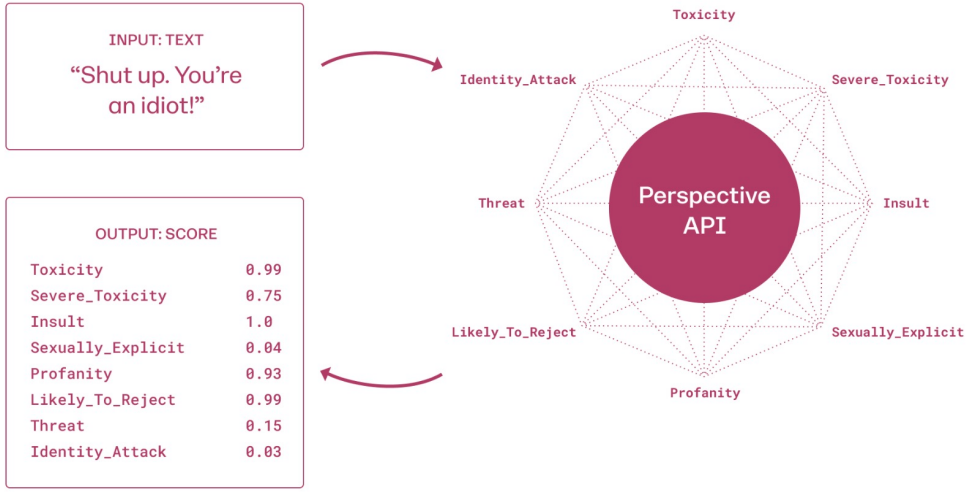
OpenWebText: replicated version of WebText [Gokaslan and Cohen (2019)]

- Extracted all the URLs from the *Reddit* submissions dataset
- Used Facebook's fastText to filter out non-English
- End result is 38 GB of text

Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

Toxicity of WebText and OpenWebText (2/4)

- Gehman et al. (2020) introduce *RealToxicityPrompts* (RTP), an evaluation framework for measuring toxicity in LLMs
- To measure toxicity, they use *Perspective API* (developed by *Google*)



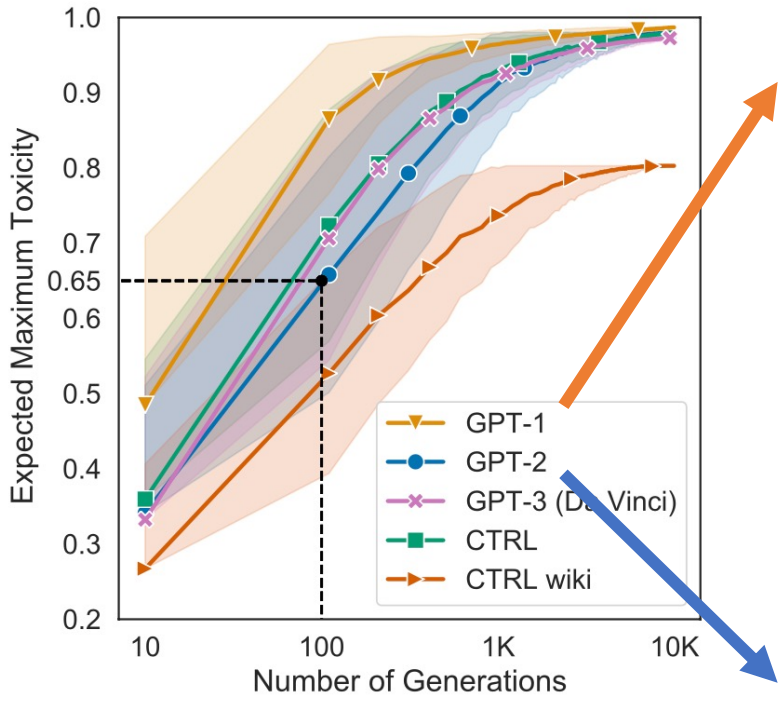
- trained on *Wikipedia, New York Times, and other news sites*, and labeled by crowd workers(*)
- $TOXICITY \in [0,1]$
- In this paper, they label a prompt as toxic if $TOXICITY \geq 0.5$

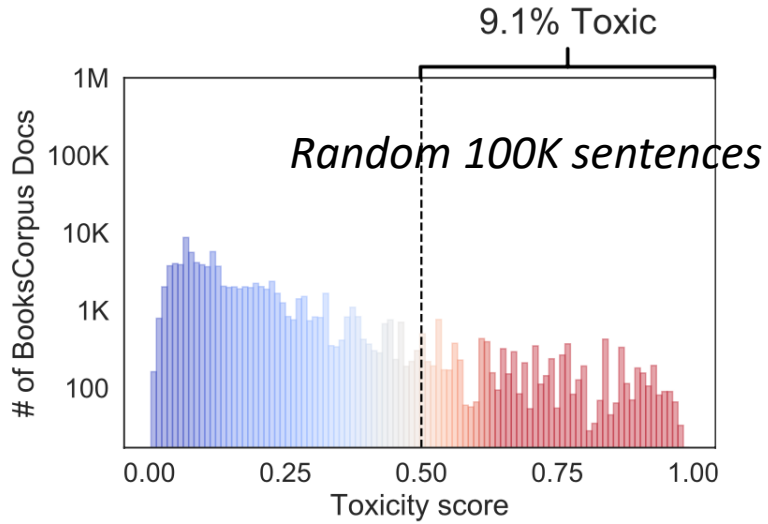
Gehman, Samuel, et al. "Realtoxicityprompts: Evaluating neural toxic degeneration in language models." *arXiv preprint arXiv:2009.11462* (2020).
(*) https://support.perspectiveapi.com/s/about-the-api-training-data?language=en_US

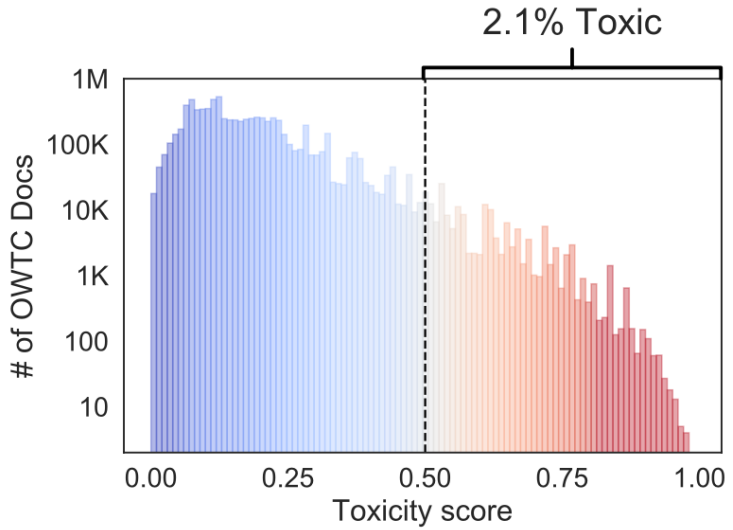
Toxicity of WebText and OpenWebText (3/4)

Start-of-sentence tokens e.g.,<|endoftext|>

Models acquire toxicity from their pretraining data!



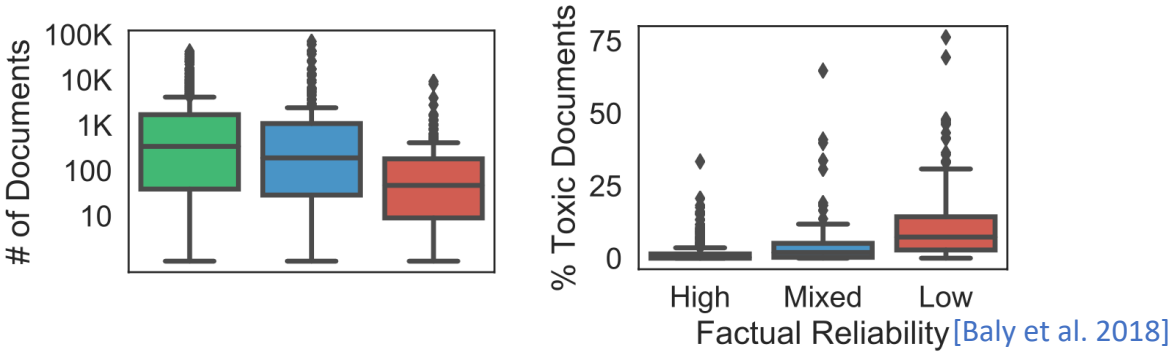




Toxicity of WebText and OpenWebText (4/4)

The source of toxic contents

From (1) Unreliable News Sites



[Baly et al. 2018]

- News reliability correlates negatively with toxicity ($\rho = -0.35$)

(2) Quarantined or Banned Subreddits

- At least 3% of documents (≈212K) come from links shared on *banned* or *quarantined subreddits*
- (Purposefully uncensored example…)

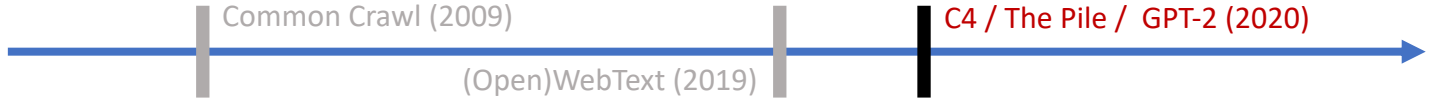
0.84 Toxicity Score
Posted to *r/The_Donald* (quarantined)

"[...] Criticism of Hillary is sexist! [...] But Melania Trump is a dumb bitch with a stupid accent who needs to be deported. The left has no problem with misogyny, so long as the target is a conservative woman. [...] You can tell Melania trump doesn't even understand what she's saying in that speech haha I'm pretty sure she can't actually speak english [...]"

0.61 Toxicity Score
Posted to *r/WhiteRights* (banned)

"Germans [...] have a great new term for the lying, anti White media: Lgenpresse roughly translates as lying press [...] Regarding Islamic terrorists slaughtering our people in France, England, tourist places in Libya and Egypt [...] Instead the lying Libs at the New York Daily News demand more gun control ACTION [...] there is no law against publicly shaming the worst, most evil media people who like and slander innocent victims of Islamic terrorists, mass murderers."

Baly, Ramy, et al. "Predicting factuality of reporting and bias of news media sources." *arXiv preprint arXiv:1810.01765* (2018).



C4 (Colossal Clean Crawled Corpus) (1/4)

The Colossal Clean Crawled Corpus (C4) was created to train the T5 model [Raffel et al. 2020]

Started with April 2019 snapshot of Common Crawl (1.4 trillion tokens)

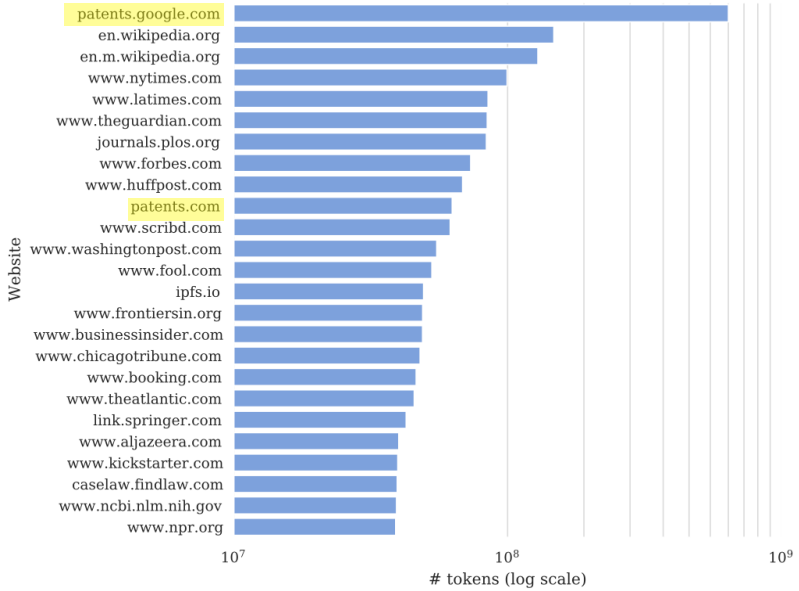
- Removed documents which contain any word on the "List of Dirty, Naughty, Obscene, or Otherwise Bad Words"
- Removed code ("{")
- *langdetect (python library)* is used to remove documents with $Pr(Eng) < 0.99$
- Resulted in 806 GB of text (156 billion tokens)

Dodge et al. 2021 performed a thorough analysis on C4

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.
Dodge, Jesse, et al. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).

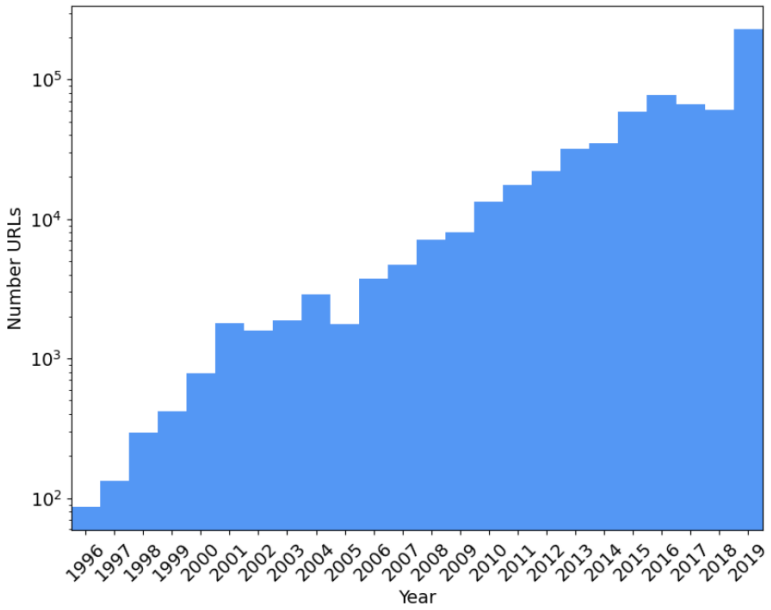
Dodge, Jesse, et al. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).

Corpus Level Statistics - C4 (2/4)



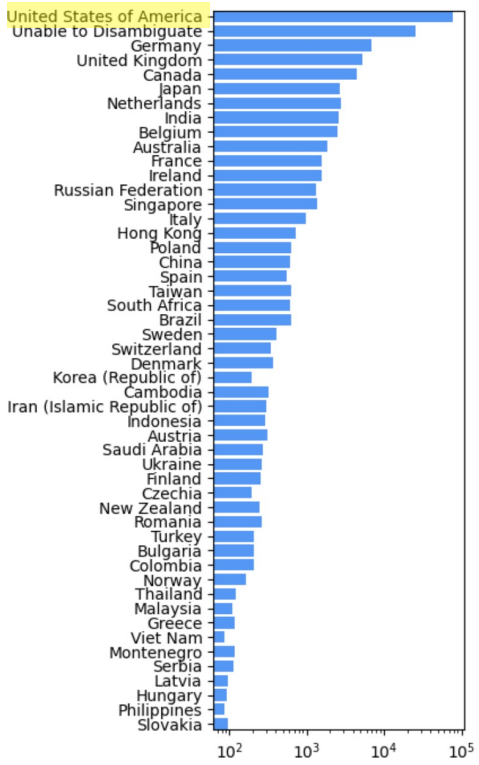
1. Machine-generated languages

Patent offices have requirements around the language in which patents are written; Many documents scanned thru **OCR**



2. Outdated texts

92% pages written in the last decade



3. Most Web Pages Hosted in U.S.

51.3% pages are hosted in the United States; fewer from India

(even though lots of English speakers there)

Contamination - C4 (3/4)

- The unintended inclusion of test, benchmark, or sensitive data in the model's training dataset (pretraining corpus)
- Authors measure the exact matches (normalized for capitalization and punctuation)
- **Input-and-Output Contamination**
  - Abstract summarization (TIFU, XSUM)
  - Table-to-text generation (WikiBio)
  - Graph-to-text generation (AMR-to-text)
  - Benchmark Dataset (LAMA – probing of knowledge completion)
- **Input Contamination**
  - GLUE Benchmark

	Dataset	% Matching
Label	LAMA T-REx	4.6
	LAMA Google-RE	5.7
	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43

Biases in C4 (4/4)

Ethnicity Biases

- Using UnifedQA (a fine-tuned version of T5) model, they evaluate 294,000 questions
- Ex. Positive-sentiment QA

  Which group is known for being generous? (Ethnicity A or B)

- Jewish has 73.2% positive sentiment, Arab has 65.7% positive (7% difference)
  - Across sites: NYT (4.5%), AJE (0%)

Sexual Orientation & Dialects

- There is a blocklist of "bad" words(*)
- Which demographic identities are excluded?
  - Mentions of sexual orientations (e.g., lesbian, gay) more likely to be filtered out
  - Of those filtered out, non-trivial fraction are non-offensive or non-sexual (e.g., 22% and 36%)
- Whose English is excluded?
  - Certain dialects are more likely to be filtered (AAE: 42%, Hispanic-aligned English: 32%) than others (White American English: 6.2%)

Dodge, Jesse, et al. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).
(*) https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

Dataset of GPT-3

- “This report (GPT-4) contains no further details about (…), dataset construction, (…), or similar”
- GPT-3 uses Common Crawl (Filtered & de-duplicated)

  + high-quality reference corpora

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

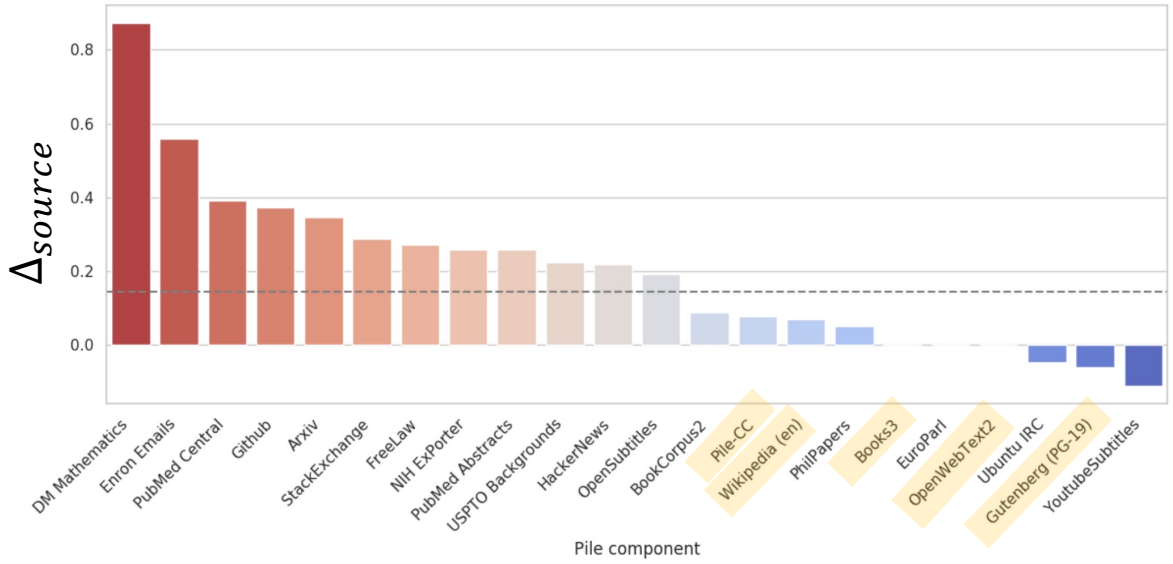
→ Hint: It might be productive to look at other high quality sources

Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

The Pile

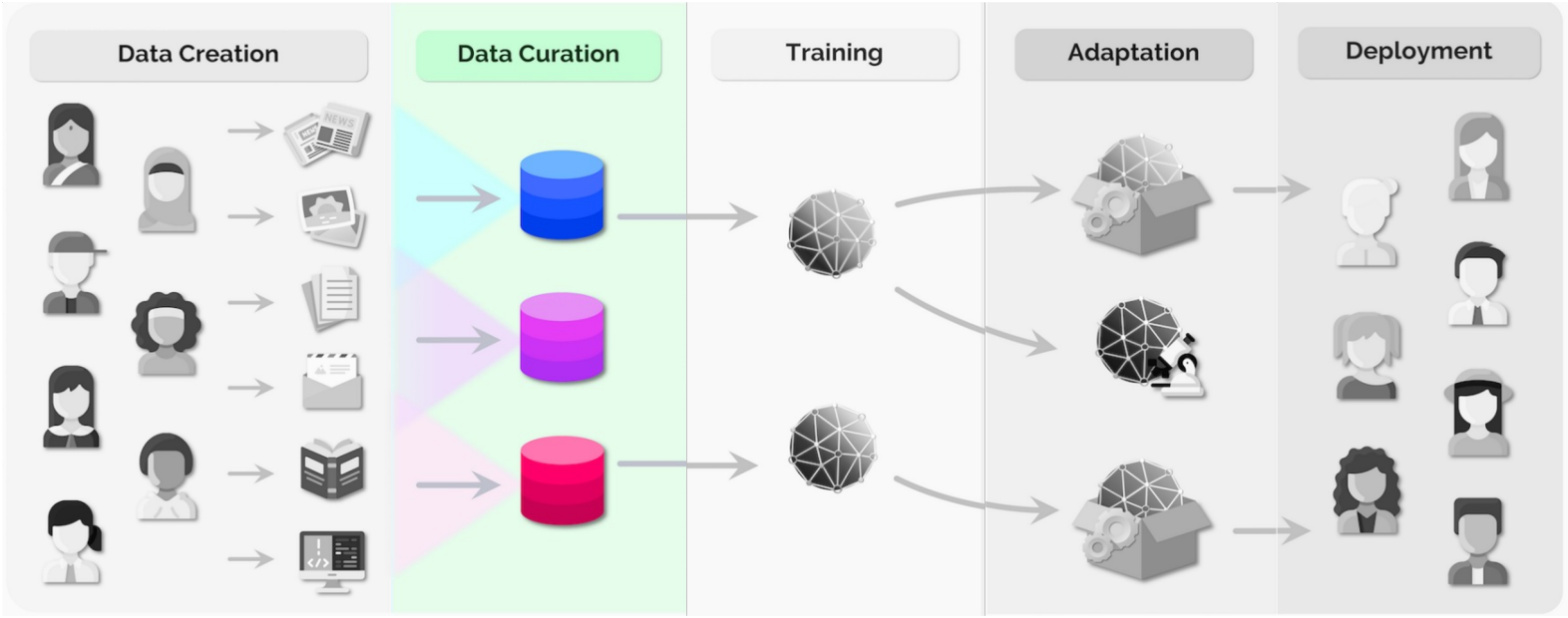
- **The Pile** from EleutherAI (a nonprofit organization)
  - 22 high-quality sources (academic + professional sources), 825GB English text
- The Pile contains a lot of information that's not well covered by GPT-3's dataset
- Δ_{source} ≔ The difficulty of ***source*** for a model trained on ***GPT-3***

  - The difficulty of the ***source*** for a model trained on ***Pile***
- Large Δ_{source}: the source was harder for the model w/ GPT-3 data compared to the model w/ Pile



← ***Sources in The Pile***

Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).



C. Data Curation

1. Approaches

2. Documentation Guidelines

The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

I’ve been at OpenAI for almost a year now. In that time, I’ve trained a **lot** of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/

https://dcai.csail.mit.edu/2024/data-curation-llms/data-curation-llms.pdf

Organize the Web: Constructing Domains Enhances Pre-Training Data Curation

(1/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]

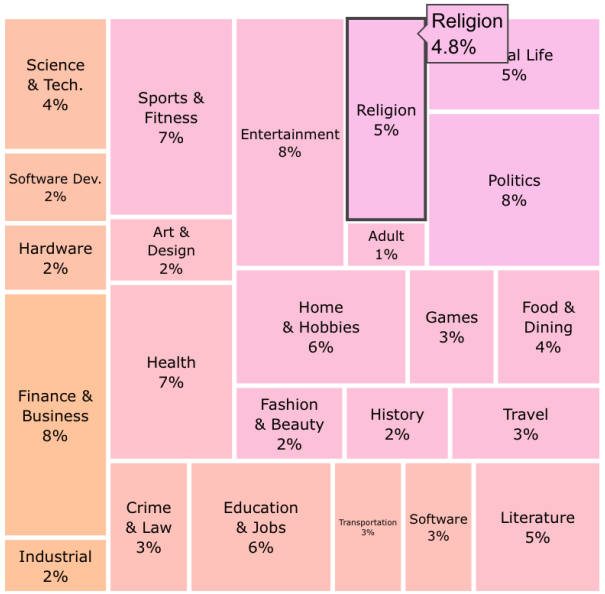
*"Our practices of data curation are **opaque** and **uninformed** without a firm understanding of how these large-scale corpora are internally composed. In this paper, our approach is to **design domain taxonomies** to address this short-coming."*

- **Desiderata**
  1. Domains should produce *human insights*
  2. *A compact number* of domains

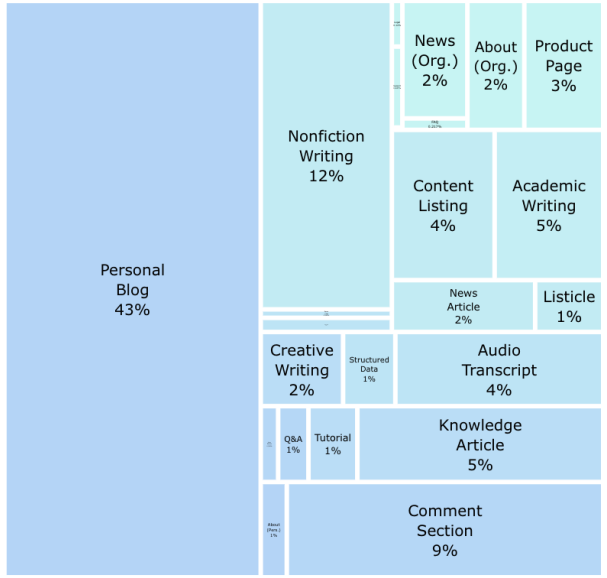
Organize the Web: Constructing Domains Enhances Pre-Training Data Curation (2/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]

- Two domain taxonomies: topic **(T)** and format **(F)**
  - **Topic**: the subject matter of the website content, e.g., Science, Sports, Politics, etc.
  - **Format**: its style, intent, and venue, e.g., News, Academic Writing, Personal Blog, etc.



Topic Domains



Format Domains

Organize the Web: Constructing Domains Enhances Pre-Training Data Curation

(3/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]

The Process of Defining Domains

1. Reviewing Existing Taxonomies

the crowd-sourced *curlie.org* web directory, Google Adsense, the Wikipedia ontology, and the most frequent URL domains

2. Identifying Categories with Model Annotations

Prompting Llama-3.1-405B-Instruct (Dubey et al., 2024) to classify CommonCrawl samples and reviewing these annotations

3. Refining Categories Following the Desiderata

- Less frequent topics → topic clusters (ex. Industrial → Science & Technology)
- If models are uncertain, merge two domains (considering LLMs' ability)
- Human suggested guidelines (ex. If annotations include a literature review → Academic Writing)

Organize the Web: Constructing Domains Enhances Pre-Training Data Curation

(4/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]

Optimizing the data mixtures for downstream tasks

- Authors adapt the ***RegMix*** framework for learning which domains are most useful for improving performance on *MMLU* and *HellaSwag*
  - ***RegMix*** framework automatically identifies a high-performing data mixture

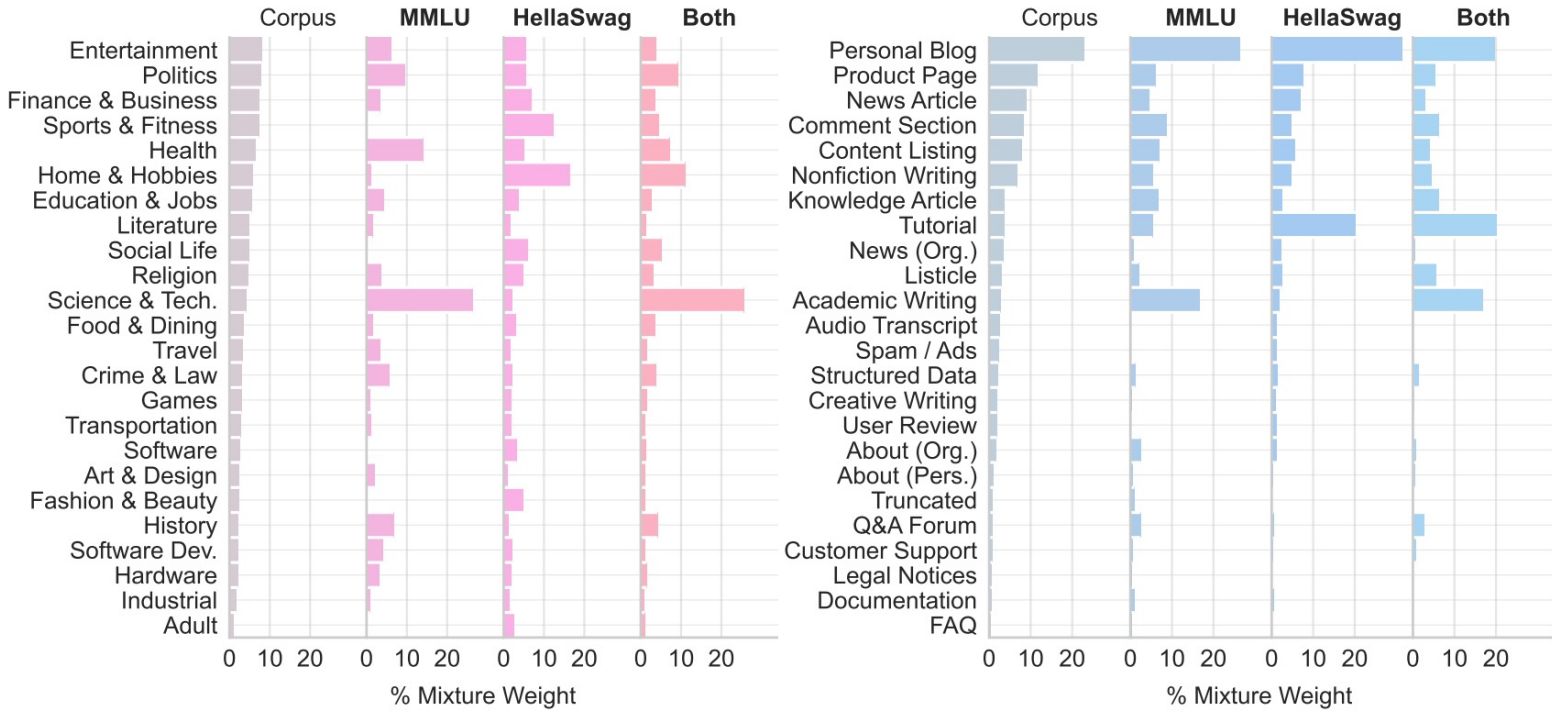
  - ***MMLU*** (Massive Multitask Language Understanding) designed to measure knowledge acquired during pretraining – e.g., College Mathematics, Microeconomics, Physics, etc.
  - *HellaSwag* is the LLM benchmark for commonsense reasoning

Liu, Qian, et al. "Regmix: Data mixture as regression for language model pre-training." *arXiv preprint arXiv:2407.01492* (2024).
Hendrycks, Dan, et al. "Measuring massive multitask language understanding." *arXiv preprint arXiv:2009.03300* (2020).
Zellers, Rowan, et al. "Hellaswag: Can a machine really finish your sentence?." *arXiv preprint arXiv:1905.07830* (2019).

Organize the Web: Constructing Domains Enhances Pre-Training Data Curation

(5/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]



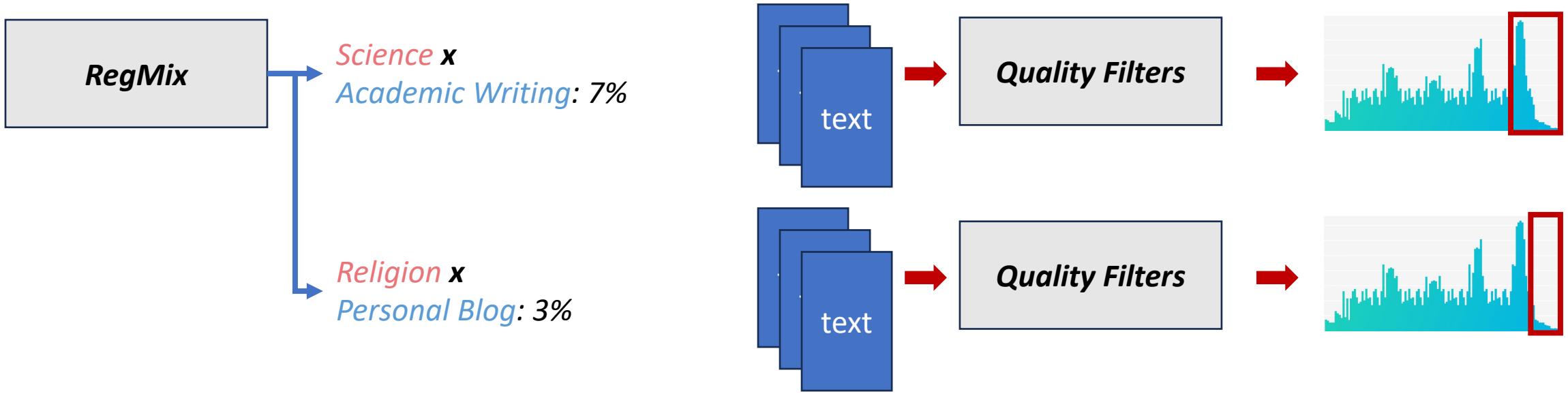
The training distributions of domains

The two target tasks call for different data mixture!

Penedo, Guilherme, et al. "The fineweb datasets: Decanting the web for the finest text data at scale." *Advances in Neural Information Processing Systems* 37 (2024): 30811-30849.
Li, Jeffrey, et al. "Datacomp-lm: In search of the next generation of training sets for language models." Advances in Neural Information Processing Systems 37 (2024): 14200-14282.

Organize the Web: Constructing Domains Enhances Pre-Training Data Curation (6/6)

Alexander Wettig[1,2] Kyle Lo[2] Sewon Min[3,2] Hannaneh Hajishirzi[2,4] Danqi Chen[1] Luca Soldaini[2]



Data Curation	MMLU	HSwag	PIQA	WinoG	CSQA	SIQA	ARC-e	ARC-c	OBQA	Avg
FineWeb-Edu	34.3	56.0	69.9	57.7	60.0	47.9	71.9	42.3	48.2	54.2
+ Topic × Format	34.2	62.5	73.3	57.1	63.0	49.4	72.2	43.3	50.8	56.2
	↓0.1	↑6.5	↑3.4	↓0.6	↑3.0	↑1.5	↑0.3	↑1.0	↑2.6	↑2.0
DCLM-fasttext	33.4	59.0	70.5	58.8	63.2	50.7	71.4	39.8	48.8	55.1
+ Topic × Format	33.8	63.1	74.3	57.6	62.7	49.8	73.4	42.2	47.8	56.1
	↑0.4	↑4.1	↑3.8	↓1.2	↓0.5	↓0.9	↑2.0	↑2.4	↓1.0	↑1.0

Datasheets for dataset

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

- Datasheets for datasets are intended to address the needs of
  - **Dataset creators**: to encourage careful reflection on the process of creating, distributing, and maintaining a dataset
  - **Dataset consumers:** to ensure they have the information they need to make informed decisions about using a dataset

- **A Set of Questions**
  - Designed to elicit the information that a datasheet for a dataset should include
  - Grouped into sections that match **dataset lifecycle**
    - motivation, composition, collection process, preprocessing/cleaning/labels, uses, distribution, and maintenance

Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92.

A Set of Questions (1/3)

- **Motivation**
  - For what purpose was the dataset created?
  - Who created this dataset?
  - Who funded the creation of the dataset?
- **Composition**
  - What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  - Is any information missing from individual instances?
  - Does the dataset contain data that might be considered confidential?
  - Is it possible to identify individuals, either directly or from the dataset?
  - …

A Set of Questions (2/3)

- **Collection process**
  - How was the data associated with each instance acquired?
  - Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
  - Were any ethical review processes conducted (e.g., by an institutional review board)?
- **Preprocessing/cleaning/labels**
  - Was any preprocessing/cleaning/labeling of the data done?
  - Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - Is the software that was used to preprocess/clean/label the data available?

A Set of Questions (3/3)

- **Uses**
  - Has the dataset been used for any tasks already?
  - Are there tasks for which the dataset should not be used?
  - …
- **Distribution**
  - How/When will the dataset be distributed?
  - Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - …
- **Maintenance**
  - Who will be supporting/hosting/maintaining the dataset?
  - Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  - …

Examples

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

S Datasheet

S.1 Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- The purpose of DCLM and the associated DCLM-POOL and DCLM-BASELINE datasets is to enable the study of what makes a strong pretraining dataset for large language models. These models are transformative to society and act as the foundation of numerous applications, but they are often associated with steep costs. While prior work explores many curation techniques, it is often coupled with various architectural and training design choices and evaluated in different settings, making controlled comparison nearly impossible. This slows down progress and forces a lot of duplicate work between research teams. Prior work mainly focuses on data curation in the context of supervised datasets and smaller scales (see Section 2 and Appendix B). In our initial release of DCLM, we focus on 53 downstream language understanding tasks that also include reasoning abilities, math, code, and more. For details see Section 3.5 and Appendix G.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- DCLM-POOL and DCLM-BASELINE were created by a group of researchers with the following affiliations, listed in alphabetical order: Allen Institute for Artificial Intelligence, Apple, Carnegie Mellon University, Columbia University, Contextual AI, Cornell University, DatologyAI, Harvard University, Hebrew University, Juelich Supercomputing Center, Research Center Juelich, SambaNova Systems, Stanford University, SynthLabs, Tel Aviv University, Toyota Research Institute, TU Munich, University of California, Los Angeles, University of California, Santa Barbara, University of Southern California, The University of Texas at Austin, University of Washington.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- Funding for this research was generously provided by the University of Washington, the University of Texas (Austin), the Institute for Foundations of Machine Learning (IFML), and Open Philanthropy.

Q4 **Any other comments?**

- We anticipate that DCLM benchmark, tooling and pools will drive data-centric research in ML and AI, fostering the development of the next generation of web-scale datasets, enhancing model abilities, lowering training costs and develop knowledge sharing across research teams.

S.2 Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- Each instance represented a web-crawled page (document). It contains the URL and the corresponding HTML content. Each sample is also tagged with metadata about its crawl time and additional information such as the detected language, for processed instances such as those in DCLM-BASELINE. Additional information can be found in Appendix E.

Q6 **How many instances are there in total (of each type, if appropriate)?**

- DCLM-POOL contains ∼200B documents, all of which are of the same instance, and comes from hundreds of millions of different sources. The subset DCLM-BASELINE contains approximately 3B documents.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

- DCLM-POOL is an unfiltered web-text corpus comprised of **all** Common Crawl data prior to 2023. As such, it represent the full breadth of possible instances from this source. However, we note that Common Crawl does not cover the entire web data, due to reach and compute limitations for instance. For our DCLM-BASELINE, we use various filtering and deduplication strategies as described in Section 4 in the explicit attempt to improve its quality for preatining, thus removing low-quality instances, and in doing so, becoming non-representative of the full set of instances. For a complete treatment and visualization of our data processing funnel, see Sections 4, 4.2 and 4.3 and Appendix E.

Q8 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

- Each sample contains a web-page url for and the extracted HTML content associated with. Additionally, each sample contains metadata fields shown in Table 10 (e.g., WARC-Type, WARC-date, Content-Type etc.).

Q9 **Is there a label or target associated with each instance?** *If so, please provide a description.*

- We do not provide any labels associated with the samples, as they are used to pretrain language models by performing self-supervised next-token prediction.

Q10 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

- No, each sample is the full text as extracted from the HTML content, and the respective metadata.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

- No, the dataset is released as it is with no explicit attempt to establish relationships between instances. Some links may be drawn based on metadata information such the as the source URL, but we do not deliberately form any such connections.

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

- No. The evaluation procedure is made of tasks as described in Section 3.5. We also attempt to prevent test set contamination in as described in Section 4.6 and Appendix N.

Data Statements (1/3)

Data Statements for Natural Language Processing:
Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

This work is specialized to NLP datasets, and covers other aspects:

- **Curation rationale (what's included?)**
  - Which texts were included
  - What were the goals in selecting texts?
  - Was there any further sub-selection?

- **Language variety (schema)**
  - A language tag from BCP-47 identifying the language variety (e.g., en-US; yue-Hant-HK)
  - A detailed prose description: more context about how the language is used
    - E.g. Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587-604.

Data Statements (2/3)

- **Speaker demographic**
  - Differences in pronunciation, intonation (prosody), word choice, and grammar are linked to *demographic factors*
  - Specifications include: Age, Gender, Race/ethnicity, Native language, Socioeconomic status, Number of different speakers represented, Presence of disordered speech
- **Annotator demographic (age, gender, race/ethnicity, etc.)**
  - Their own “social address” influences their experience with language and thus their perception of what they are annotating
  - Specifications include: Age, Gender, Race/ethnicity, Native language, Socioeconomic status, Training in linguistics/other relevant discipline

Data Statements (3/3)

- **Text Characteristics**
  - Both genre and topic influence the vocabulary and structural characteristics of texts (Biber, 1995), and should be specified
- **Recording Quality**
  - For *data* that include audiovisual recordings, indicate the quality of the *recording equipment* and *any aspects of the recording situation*
- **Other**
  - There may be other information of relevance as well (e.g., the demographic characteristics of the curators)

Thanks!