CS 329X: Human Centered LLMs

Data, Data, Data

Rose E. Wang

Announcements

- Deadlines:
    - Project proposal: Oct 10th, Thursday
        - Submission portal open on Canvas
    - HW1: Oct 16th, next week Wednesday
        - If you're waiting on compute units, try sharing your notebook to a different Google account
        - If you get CUDA out of memory, make your batch size and model smaller
        - With the DPO trainer, you need to use dpo_trainer.save_model or the model state will not properly be captured when saving away
        - You will have 270 annotations, not 180 (ignore the old assertion). Upload our preferences data as a CSV file.
- Course-level report logistics

Learning goals

- Demystify the LLM training process wrt data. Training ChatGPT is **not** just running

```
python train.py --dataset=the_dataset
```

- Lots of choices for "`the_dataset`"
- Human-made and have real-world implications!

Outline

- Why talk about data?
- What is data?
- How to process data?
    - Colab!
- What impact does data processing have on downstream applications?
- Hot-take discussion (20 minutes)

Outline

- **Why talk about data?**
- What is data?
- How to process data?
  - Colab!
- What impact does data processing have on downstream applications?
- Hot-take discussion (20 minutes)

Why are we talking about data?

Data is the most important thing in developing human-centered LLMs.

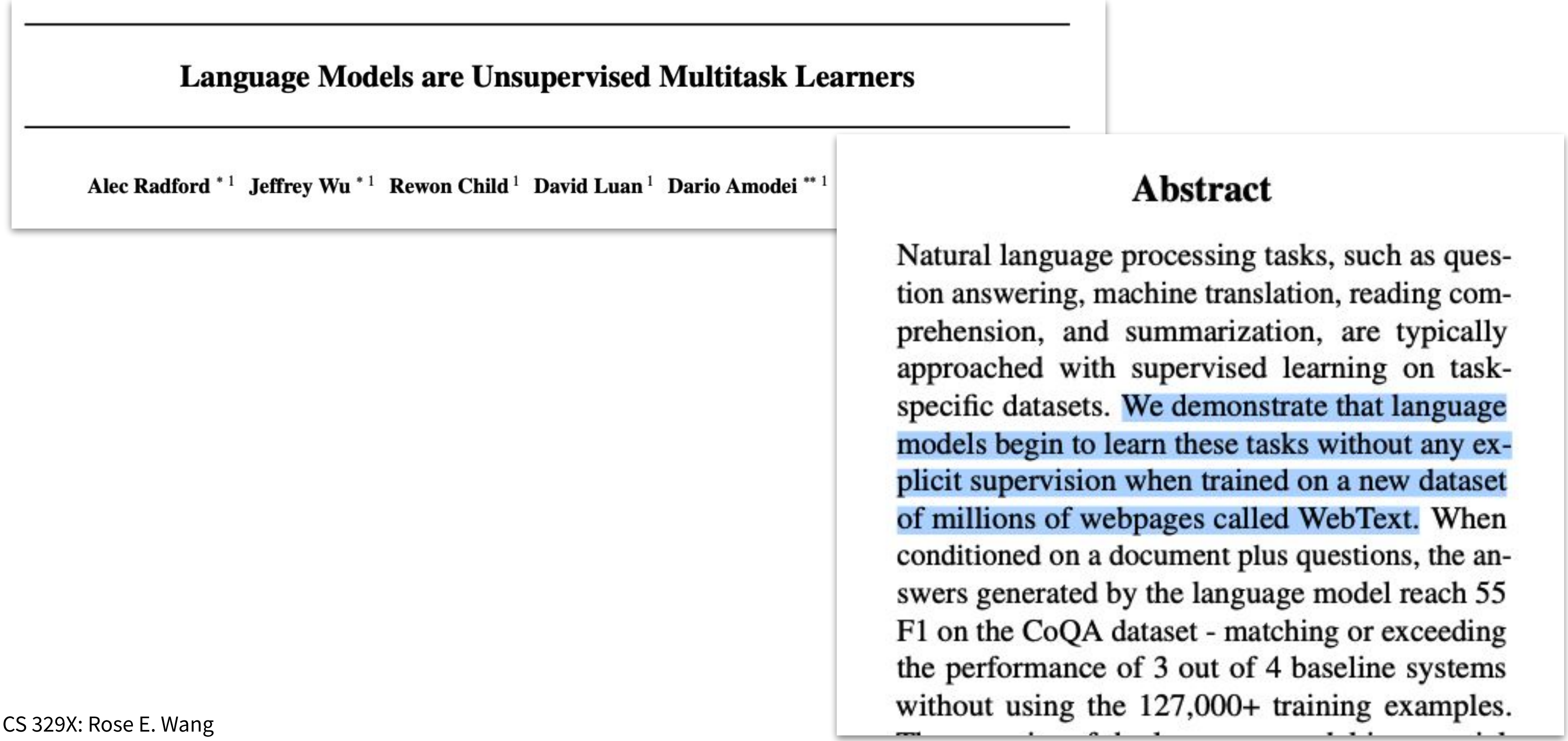
Data in the LLM pipeline

Pre-training

Fine-tuning
(e.g., preferences or domain-specific)

Evaluation / Human-Centered Implications

Why are we talking about data?



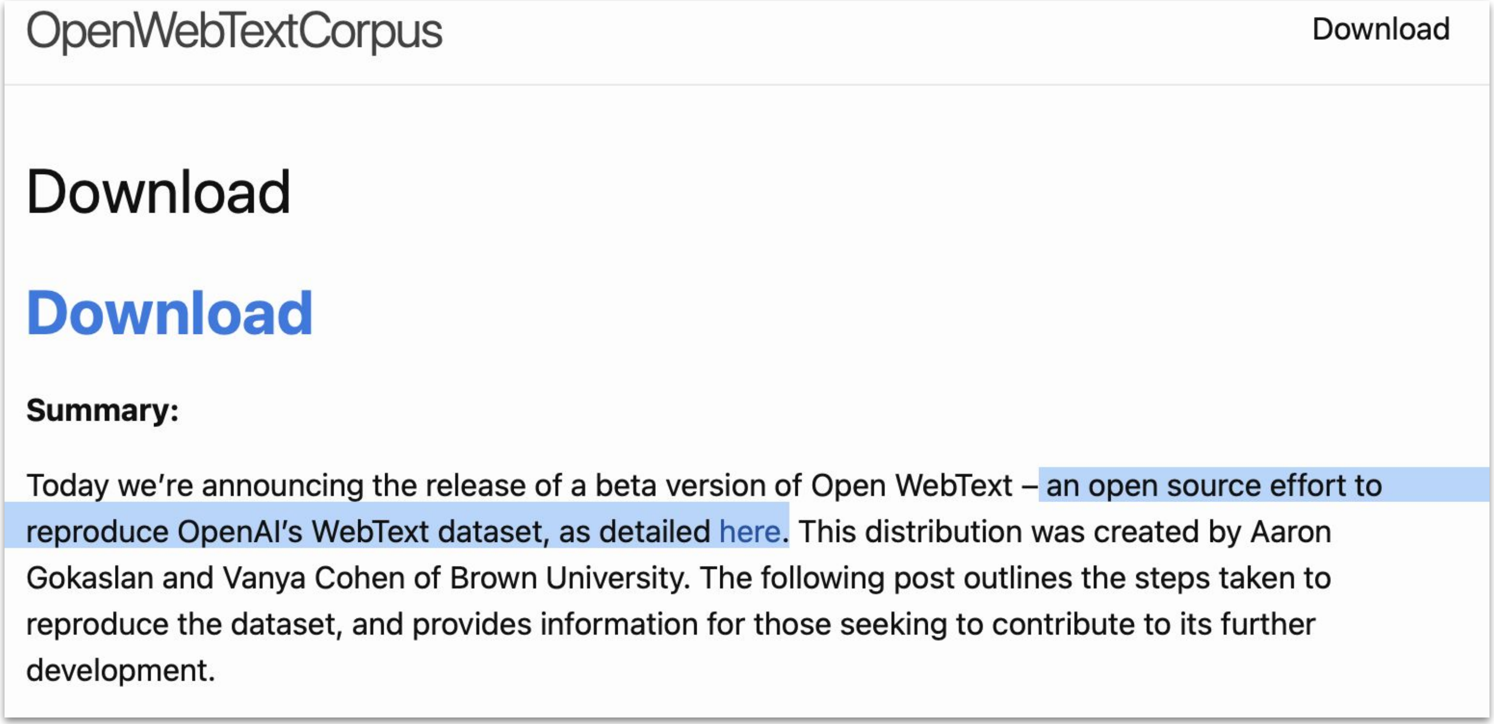
Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples.

Why are we talking about data?



OpenWebTextCorpus

Download

Download

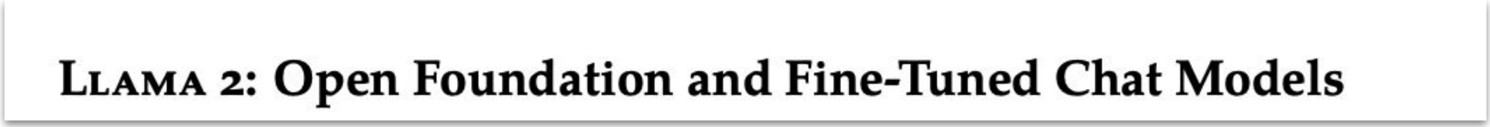
Download

Summary:

Today we're announcing the release of a beta version of Open WebText – an open source effort to reproduce OpenAI's WebText dataset, as detailed here. This distribution was created by Aaron Gokaslan and Vanya Cohen of Brown University. The following post outlines the steps taken to reproduce the dataset, and provides information for those seeking to contribute to its further development.

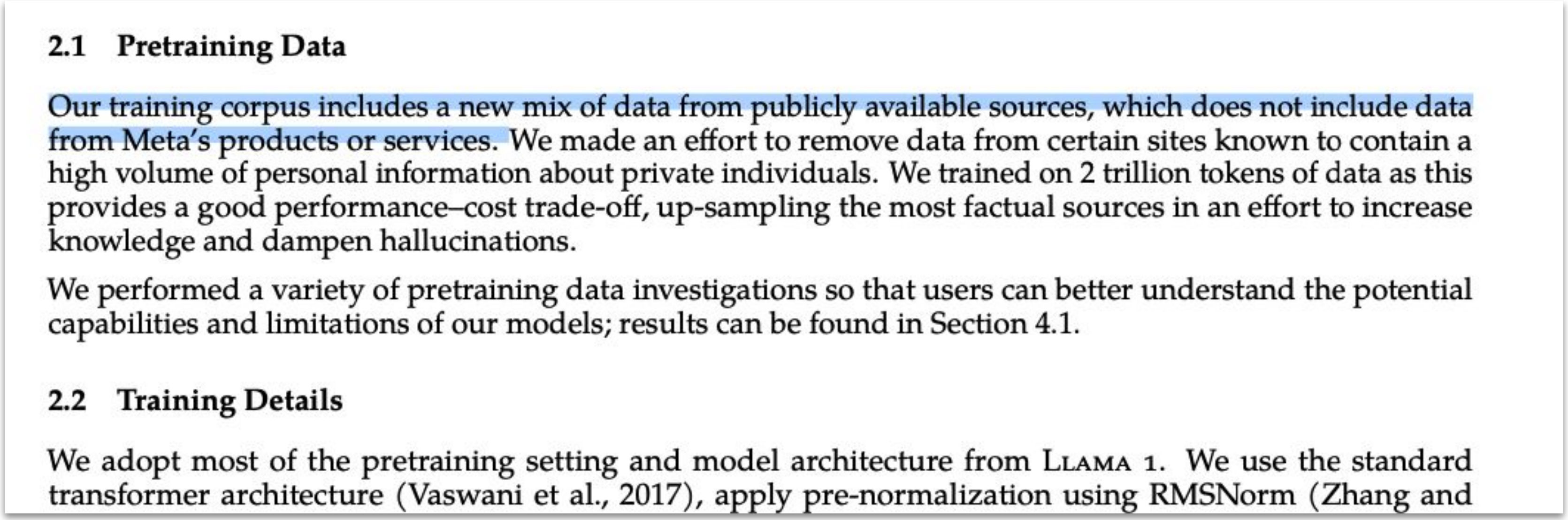
https://skylion007.github.io/OpenWebTextCorpus/

Even for open-sourced models, we have closed-sourced data!



Llama 2: Open Foundation and Fine-Tuned Chat Models

GenAI, Meta



2.1 Pretraining Data

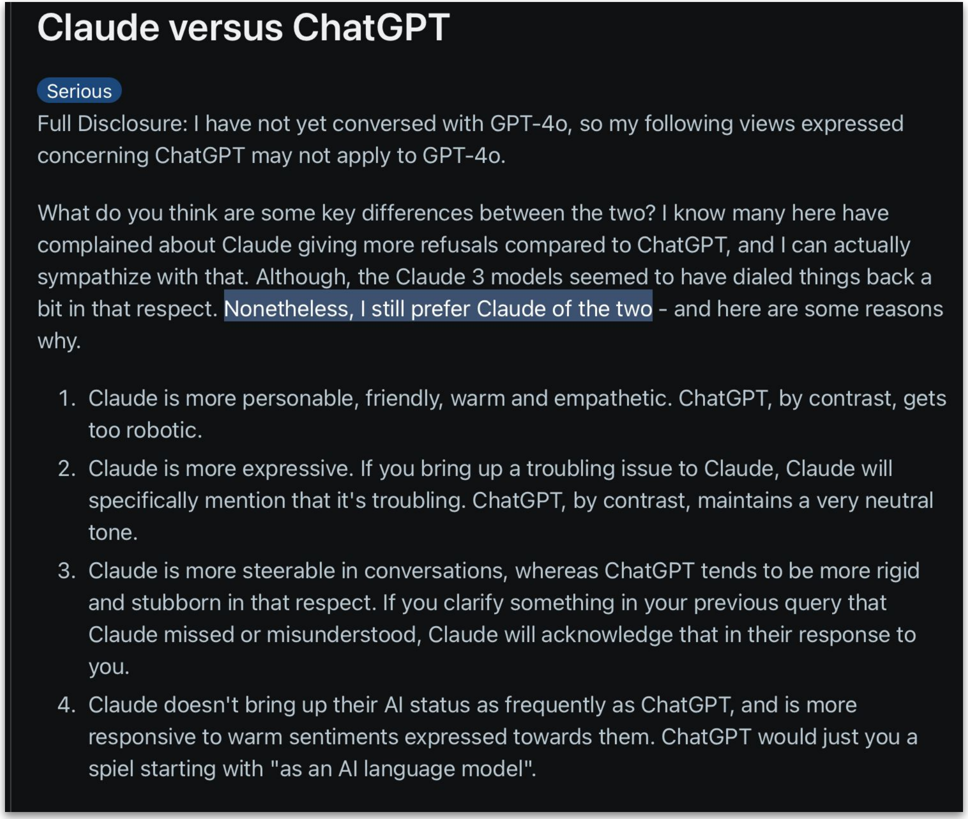
Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.

We performed a variety of pretraining data investigations so that users can better understand the potential capabilities and limitations of our models; results can be found in Section 4.1.

2.2 Training Details

We adopt most of the pretraining setting and model architecture from Llama 1. We use the standard transformer architecture (Vaswani et al., 2017), apply pre-normalization using RMSNorm (Zhang and

Reason: Competitive LLM market



Claude versus ChatGPT

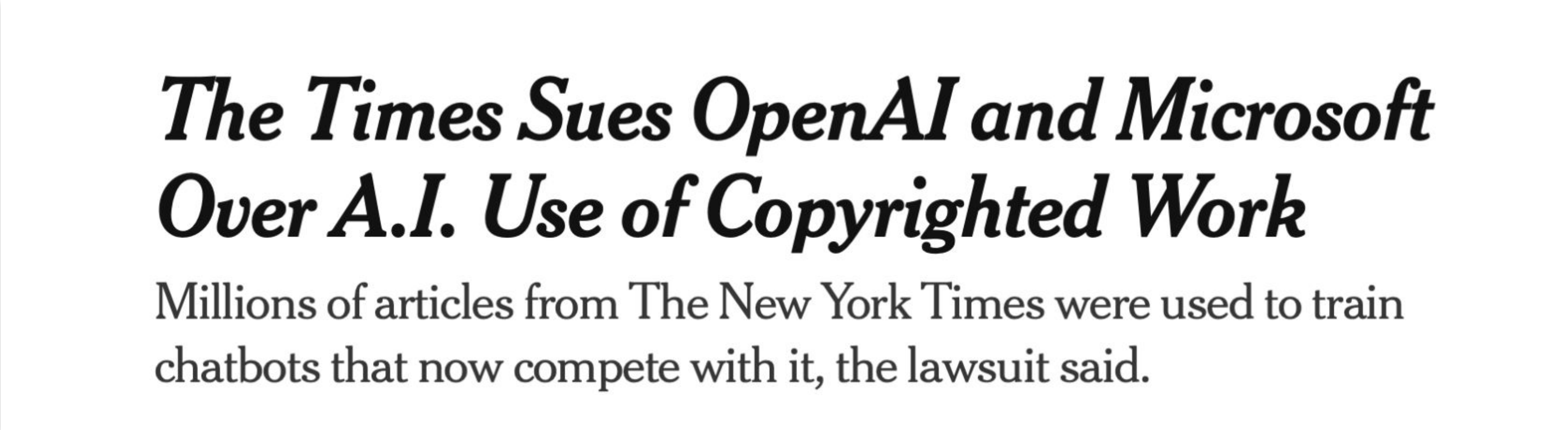
Serious

Full Disclosure: I have not yet conversed with GPT-4o, so my following views expressed concerning ChatGPT may not apply to GPT-4o.

What do you think are some key differences between the two? I know many here have complained about Claude giving more refusals compared to ChatGPT, and I can actually sympathize with that. Although, the Claude 3 models seemed to have dialed things back a bit in that respect. Nonetheless, I still prefer Claude of the two - and here are some reasons why.

1. Claude is more personable, friendly, warm and empathetic. ChatGPT, by contrast, gets too robotic.
2. Claude is more expressive. If you bring up a troubling issue to Claude, Claude will specifically mention that it's troubling. ChatGPT, by contrast, maintains a very neutral tone.
3. Claude is more steerable in conversations, whereas ChatGPT tends to be more rigid and stubborn in that respect. If you clarify something in your previous query that Claude missed or misunderstood, Claude will acknowledge that in their response to you.
4. Claude doesn't bring up their AI status as frequently as ChatGPT, and is more responsive to warm sentiments expressed towards them. ChatGPT would just you a spiel starting with "as an AI language model".

Legal reasons: e.g., copyright and liability

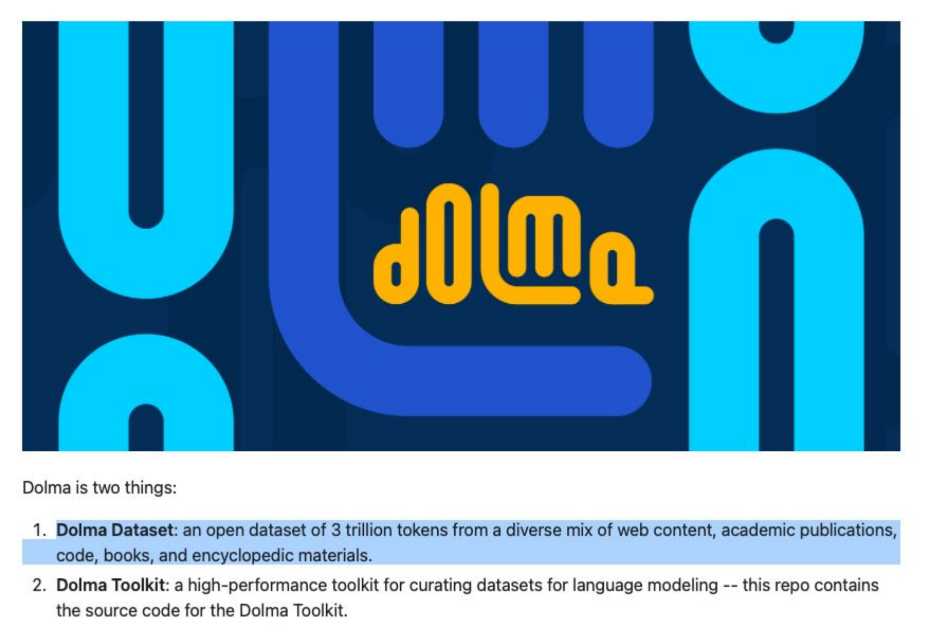


The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Implication: We can’t really make claims on model perf.

We care about generalization. But what if there’s train-test contamination?



Dolma is two things:

1. **Dolma Dataset**: an open dataset of 3 trillion tokens from a diverse mix of web content, academic publications, code, books, and encyclopedic materials.
2. **Dolma Toolkit**: a high-performance toolkit for curating datasets for language modeling -- this repo contains the source code for the Dolma Toolkit.

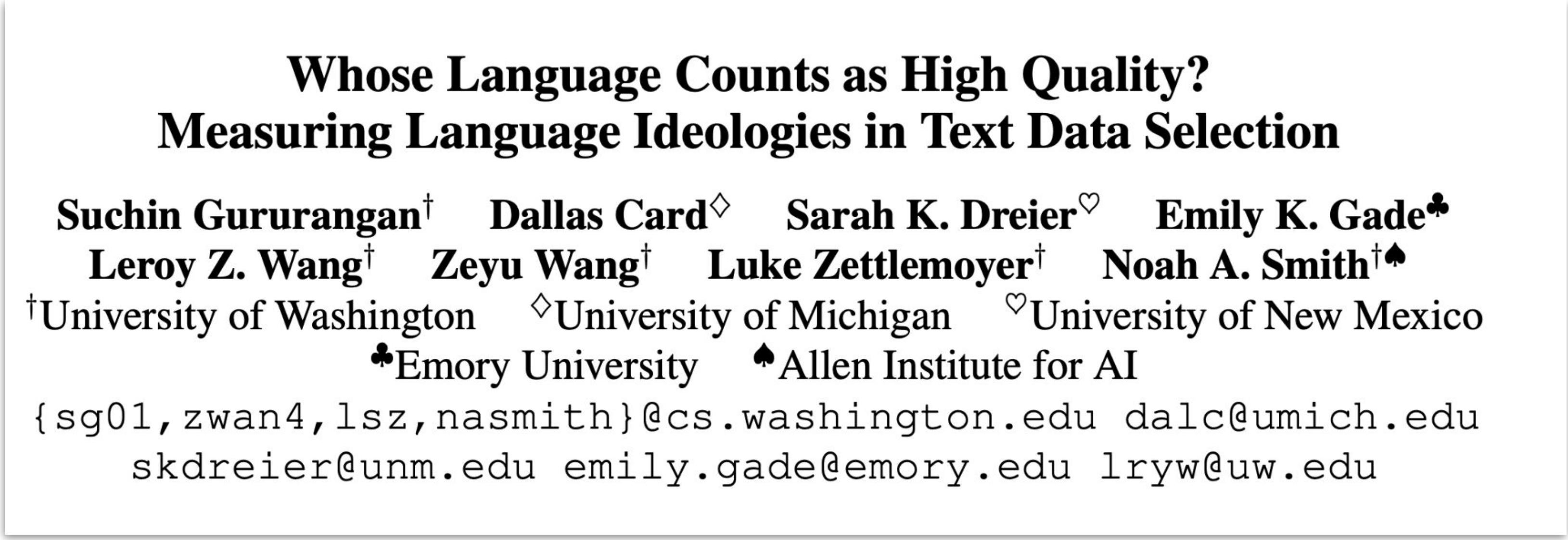
Language models

Open Language Model: OLMo

Exciting open-source efforts from AI2

Implication: Who/what are we including/excluding?

e.g., today’s readings:



Whose Language Counts as High Quality?
Measuring Language Ideologies in Text Data Selection

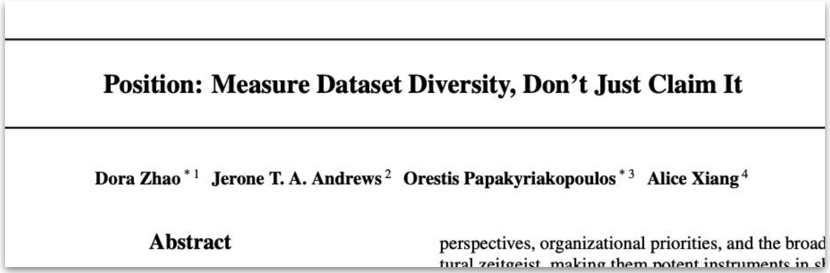
Suchin Gururangan[†] **Dallas Card**[◊] **Sarah K. Dreier**[♡] **Emily K. Gade**[♣]
Leroy Z. Wang[†] **Zeyu Wang**[†] **Luke Zettlemoyer**[†] **Noah A. Smith**[†♠]
[†]University of Washington [◊]University of Michigan [♡]University of New Mexico
[♣]Emory University [♠]Allen Institute for AI
{sg01,zwan4,lsz,nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

Today’s lecture: Choices in pre-training that have downstream implications.

Pre-training

Fine-tuning
(e.g., preferences or domain-specific)

Evaluation / Human-Centered Implications



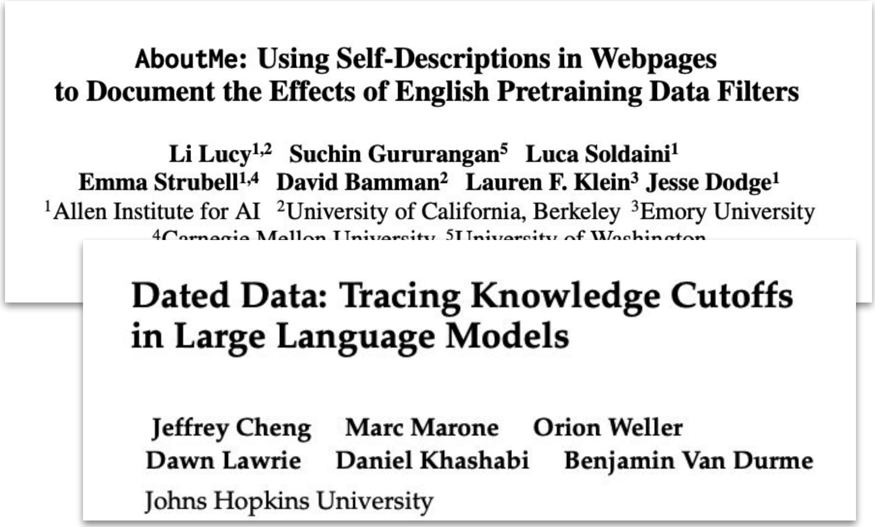
Position: Measure Dataset Diversity, Don’t Just Claim It

Dora Zhao[*1] Jerone T. A. Andrews[2] Orestis Papakyriakopoulos[*3] Alice Xiang[4]

Abstract

perspectives, organizational priorities, and the broad
tural zeitgeist, making them potent instruments in sh

→



AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters

Li Lucy[1,2] Suchin Gururangan[5] Luca Soldaini[1]
Emma Strubell[1,4] David Bamman[2] Lauren F. Klein[3] Jesse Dodge[1]
[1]Allen Institute for AI [2]University of California, Berkeley [3]Emory University
[4]Carnegie Mellon University [5]University of Washington

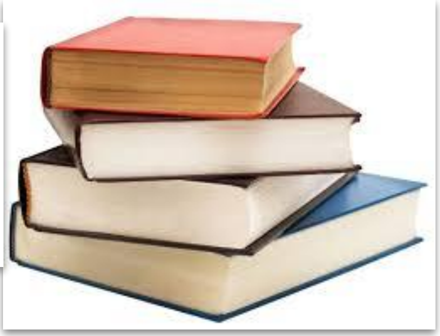
Dated Data: Tracing Knowledge Cutoffs in Large Language Models

Jeffrey Cheng Marc Marone Orion Weller
Dawn Lawrie Daniel Khashabi Benjamin Van Durme
Johns Hopkins University

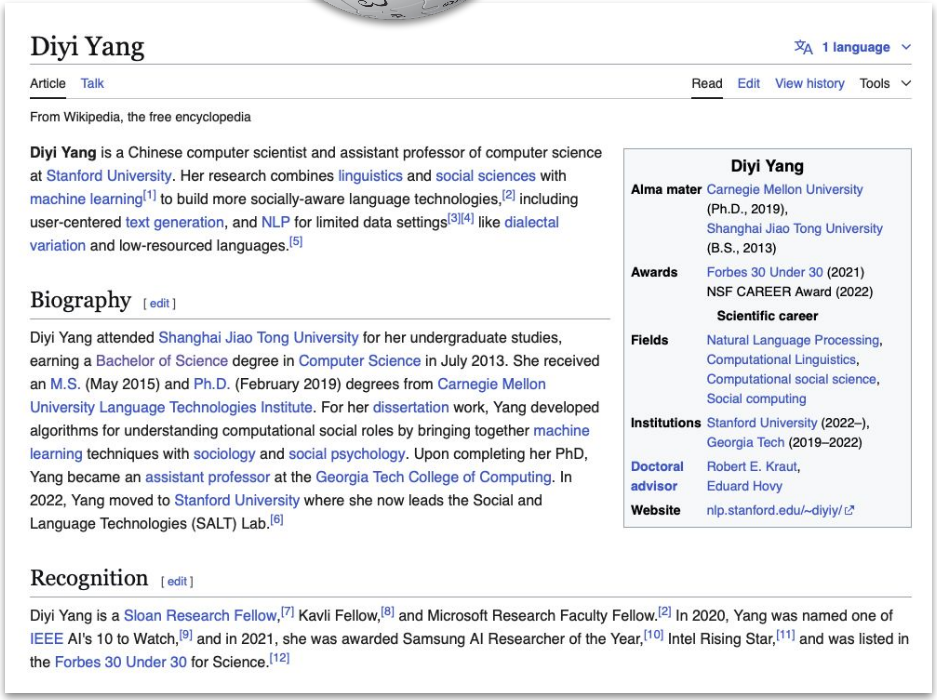
Outline

- Why talk about data?
- **What is data?**
- How to process data?
    - Colab!
- What impact does data processing have on downstream applications?
- Hot-take discussion (20 minutes)

What is data?



What is data?



Diyi Yang

From Wikipedia, the free encyclopedia

Diyi Yang is a Chinese computer scientist and assistant professor of computer science at Stanford University. Her research combines linguistics and social sciences with machine learning[1] to build more socially-aware language technologies,[2] including user-centered text generation, and NLP for limited data settings[3][4] like dialectal variation and low-resourced languages.[5]

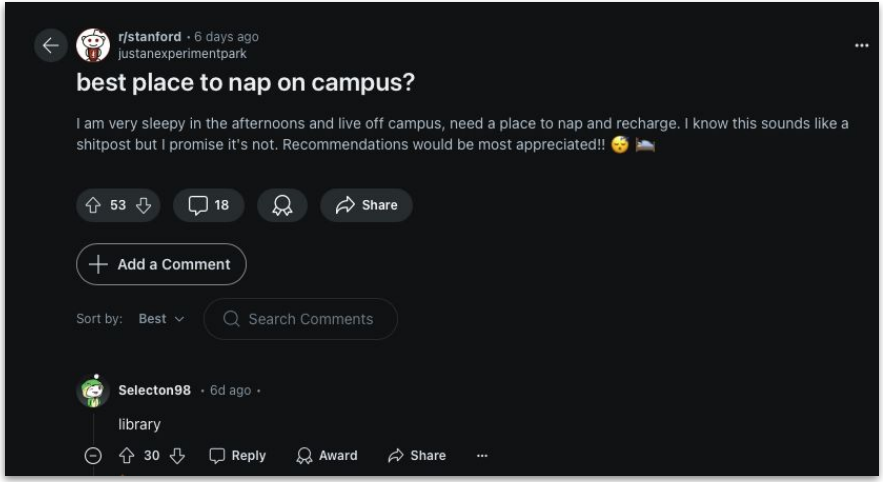
Diyi Yang	
Alma mater	Carnegie Mellon University (Ph.D., 2019), Shanghai Jiao Tong University (B.S., 2013)
Awards	Forbes 30 Under 30 (2021) NSF CAREER Award (2022)
Scientific career	
Fields	Natural Language Processing, Computational Linguistics, Computational social science, Social computing
Institutions	Stanford University (2022–), Georgia Tech (2019–2022)
Doctoral advisor	Robert E. Kraut, Eduard Hovy
Website	nlp.stanford.edu/~diyiy/

Biography [edit]

Diyi Yang attended Shanghai Jiao Tong University for her undergraduate studies, earning a Bachelor of Science degree in Computer Science in July 2013. She received an M.S. (May 2015) and Ph.D. (February 2019) degrees from Carnegie Mellon University Language Technologies Institute. For her dissertation work, Yang developed algorithms for understanding computational social roles by bringing together machine learning techniques with sociology and social psychology. Upon completing her PhD, Yang became an assistant professor at the Georgia Tech College of Computing. In 2022, Yang moved to Stanford University where she now leads the Social and Language Technologies (SALT) Lab.[6]

Recognition [edit]

Diyi Yang is a Sloan Research Fellow,[7] Kavli Fellow,[8] and Microsoft Research Faculty Fellow.[2] In 2020, Yang was named one of IEEE AI's 10 to Watch,[9] and in 2021, she was awarded Samsung AI Researcher of the Year,[10] Intel Rising Star,[11] and was listed in the Forbes 30 Under 30 for Science.[12]



r/stanford • 6 days ago
justanexperimentpark

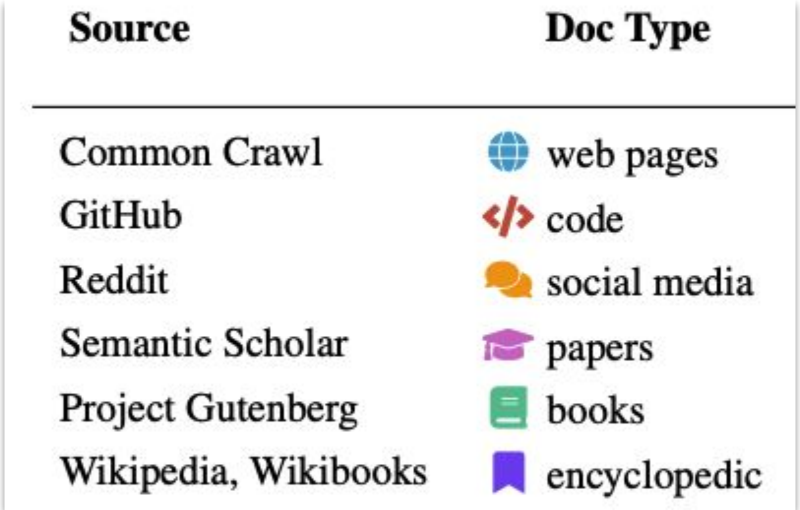
best place to nap on campus?

I am very sleepy in the afternoons and live off campus, need a place to nap and recharge. I know this sounds like a shitpost but I promise it's not. Recommendations would be most appreciated!! 😴 🛌

Selecton98 • 6d ago •

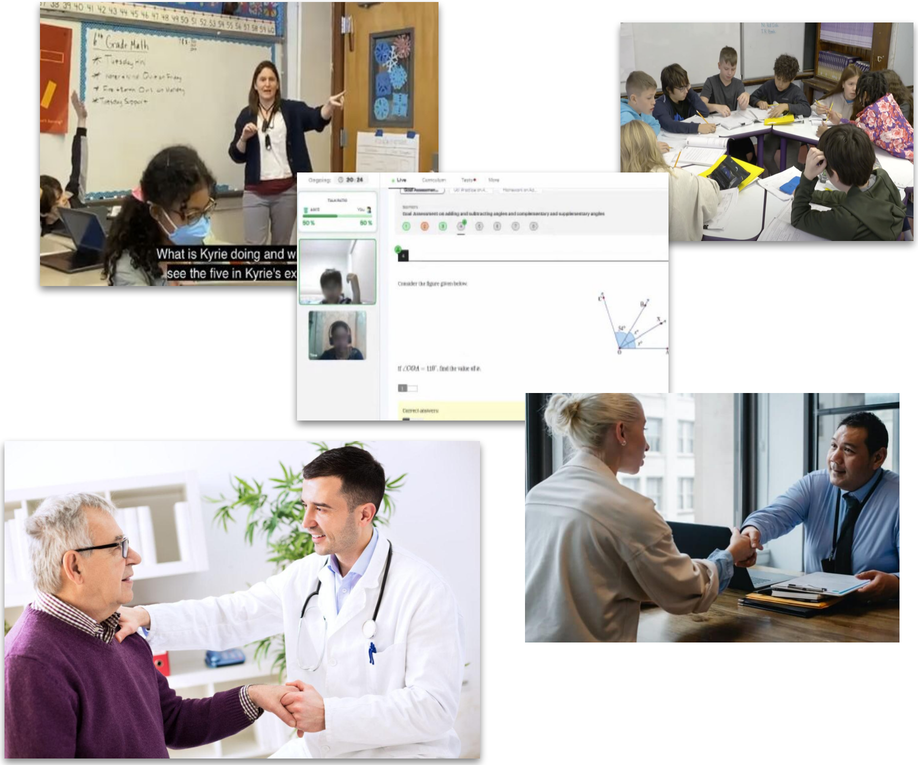
library

What is ***not*** considered data?



Source	Doc Type
Common Crawl	web pages
GitHub	code
Reddit	social media
Semantic Scholar	papers
Project Gutenberg	books
Wikipedia, Wikibooks	encyclopedic

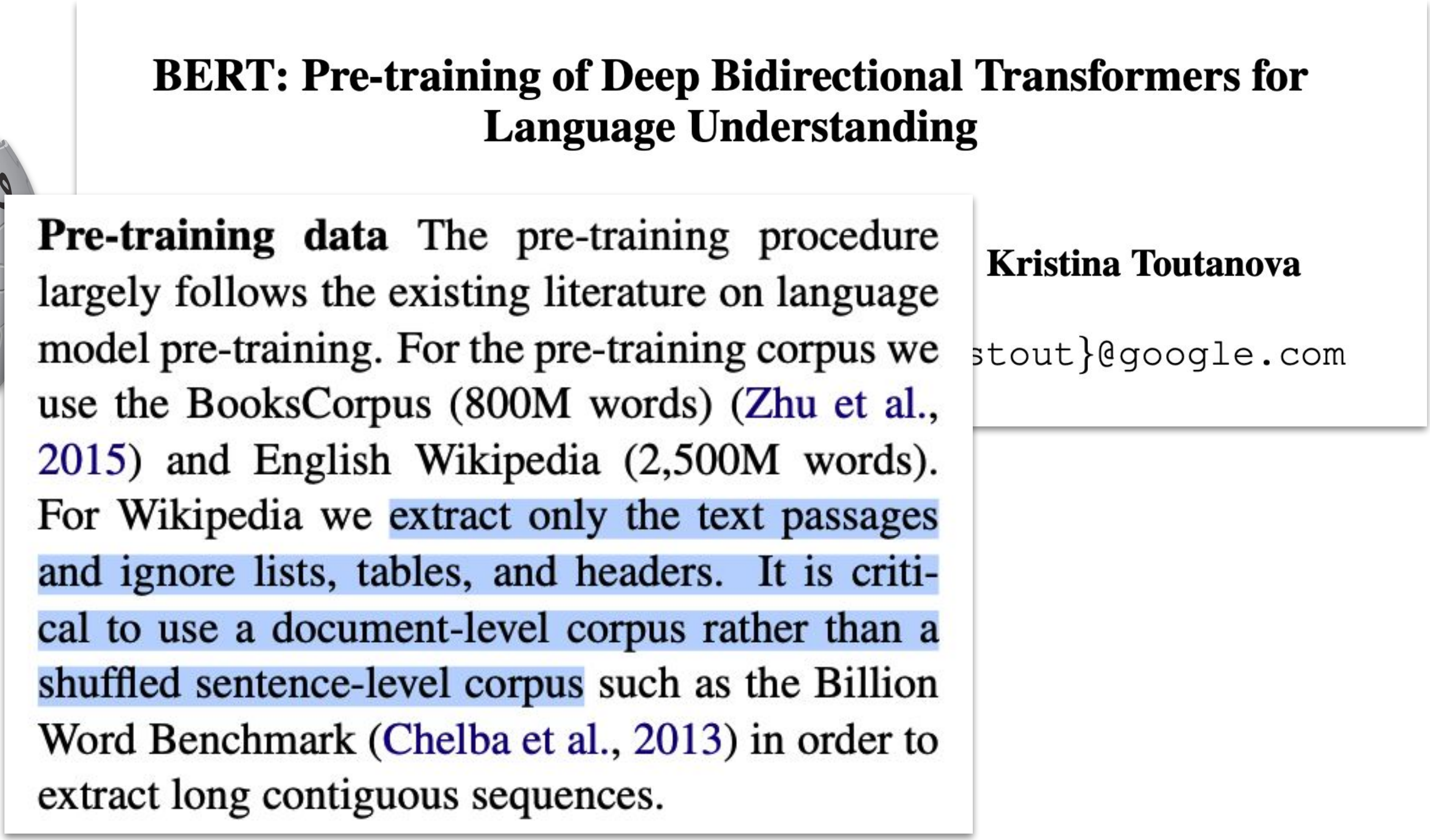
≠



What is data? Running through some examples…

Raw internet data + choices around data cleaning

What is data? Running through some examples…



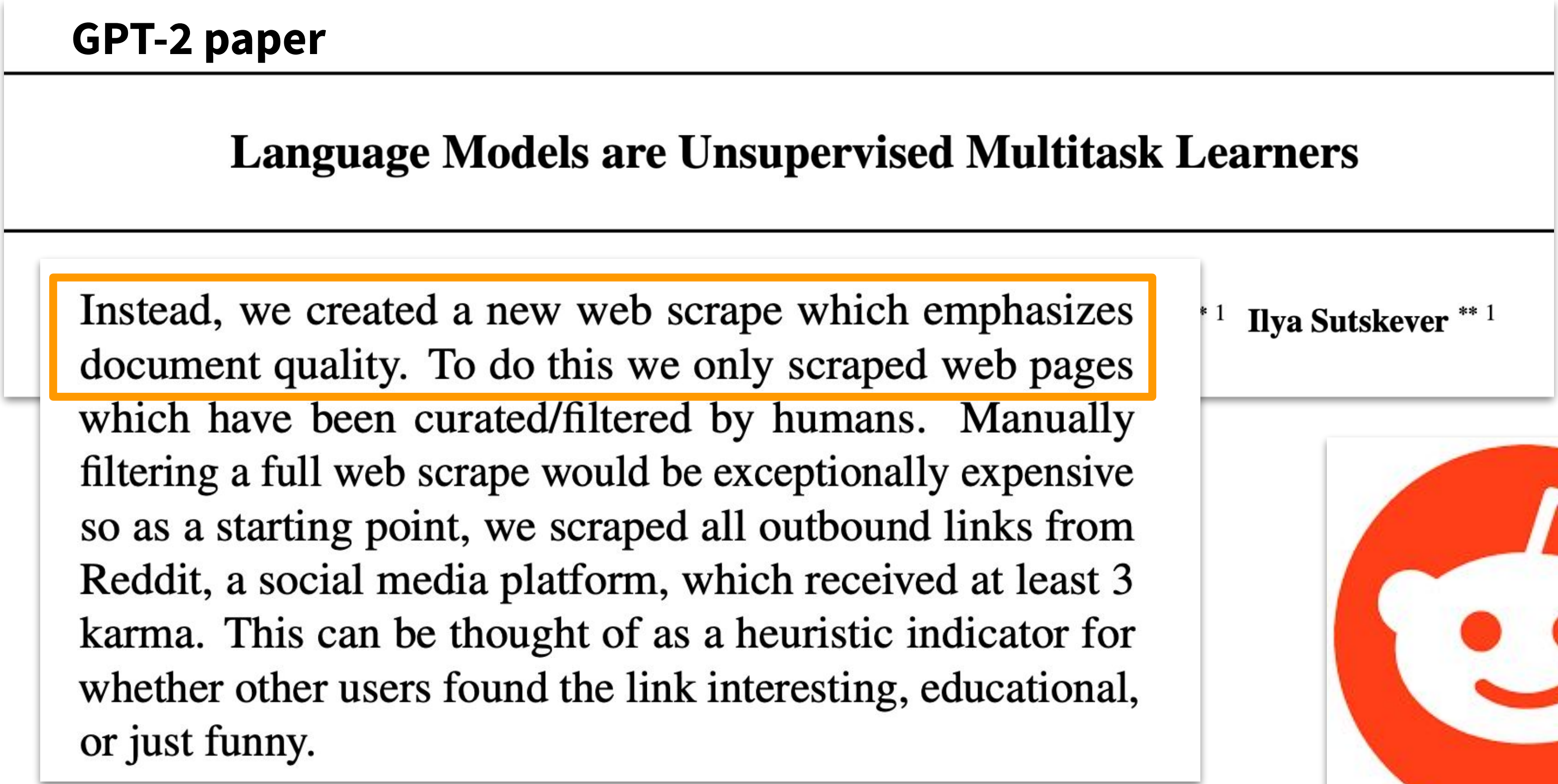
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Kristina Toutanova

stout}@google.com

Pre-training data The pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus we use the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia we extract only the text passages and ignore lists, tables, and headers. It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the Billion Word Benchmark (Chelba et al., 2013) in order to extract long contiguous sequences.

What is data? Running through some examples…



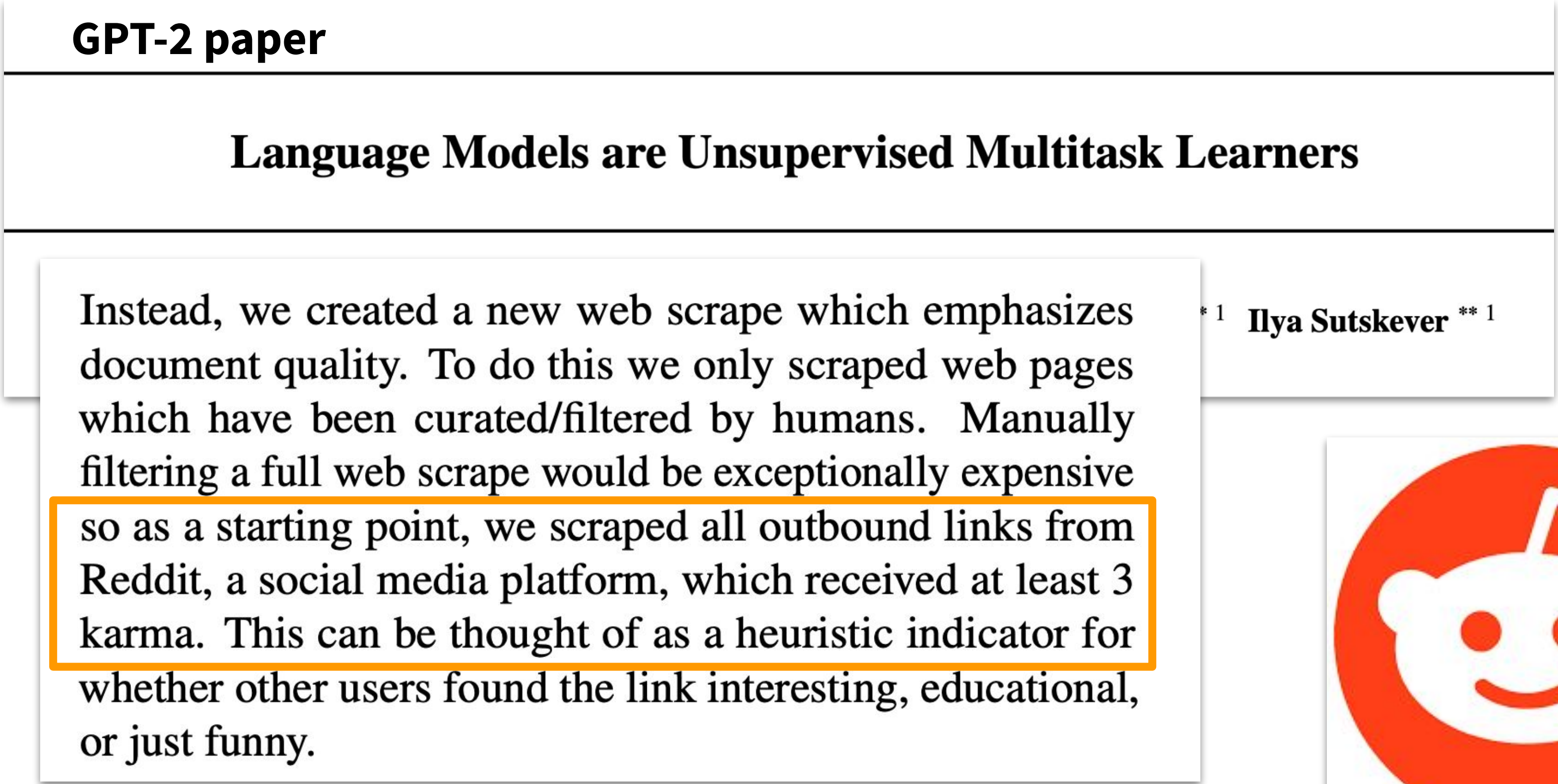
GPT-2 paper

Language Models are Unsupervised Multitask Learners

* 1 **Ilya Sutskever** ** 1

Instead, we created a new web scrape which emphasizes document quality. To do this we only scraped web pages which have been curated/filtered by humans. Manually filtering a full web scrape would be exceptionally expensive so as a starting point, we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

What is data? Running through some examples…



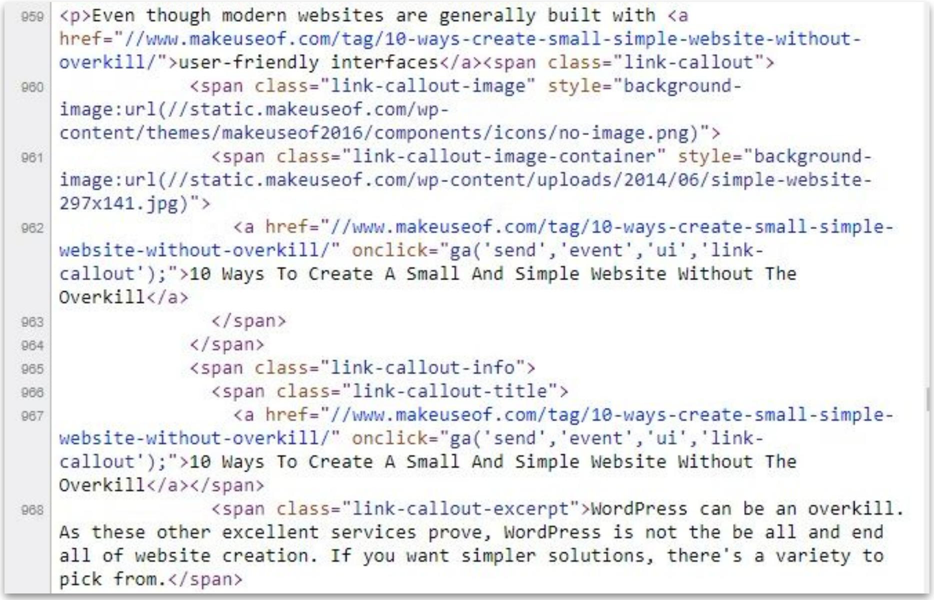
GPT-2 paper

Language Models are Unsupervised Multitask Learners

* 1 **Ilya Sutskever** ** 1

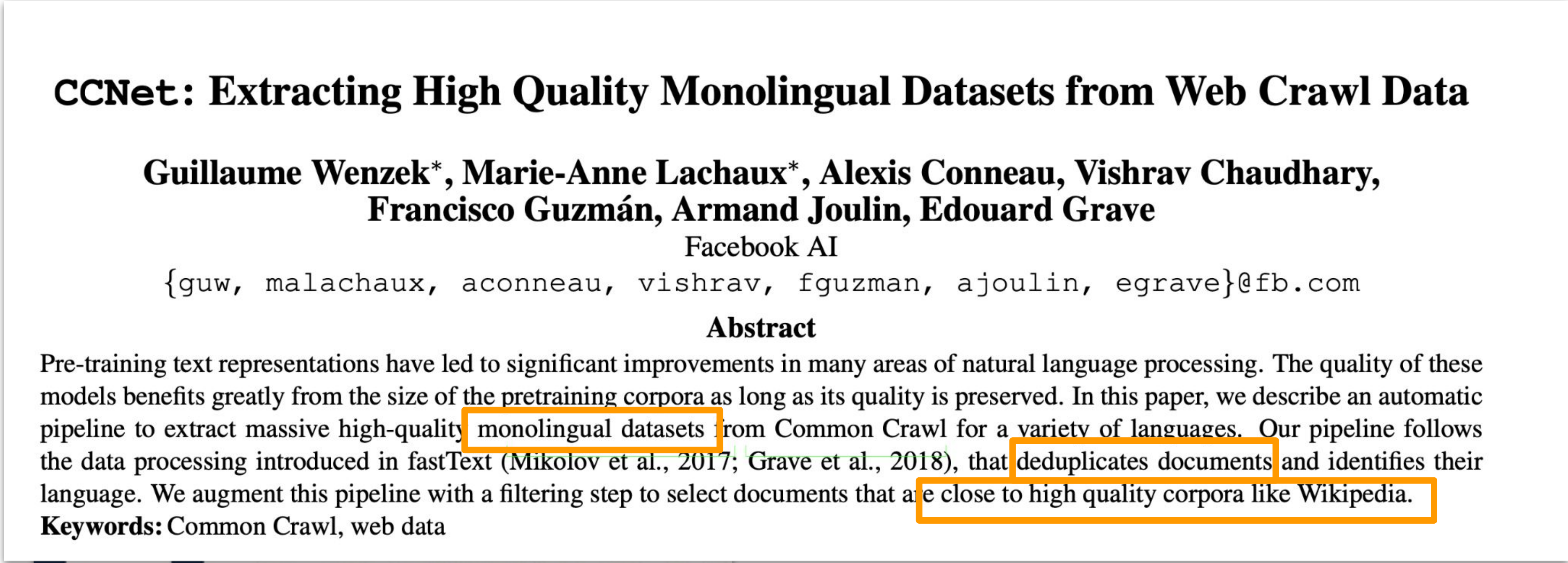
Instead, we created a new web scrape which emphasizes document quality. To do this we only scraped web pages which have been curated/filtered by humans. Manually filtering a full web scrape would be exceptionally expensive so as a starting point, we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

What is data? Running through some examples…



```
<p>Even though modern websites are generally built with <a
href="//www.makeuseof.com/tag/10-ways-create-small-simple-website-without-
overkill/">user-friendly interfaces</a><span class="link-callout">
            <span class="link-callout-image" style="background-
image:url(//static.makeuseof.com/wp-
content/themes/makeuseof2016/components/icons/no-image.png)">
              <span class="link-callout-image-container" style="background-
image:url(//static.makeuseof.com/wp-content/uploads/2014/06/simple-website-
297x141.jpg)">
                <a href="//www.makeuseof.com/tag/10-ways-create-small-simple-
website-without-overkill/" onclick="ga('send','event','ui','link-
callout');">10 Ways To Create A Small And Simple Website Without The
Overkill</a>
              </span>
            </span>
            <span class="link-callout-info">
              <span class="link-callout-title">
                <a href="//www.makeuseof.com/tag/10-ways-create-small-simple-
website-without-overkill/" onclick="ga('send','event','ui','link-
callout');">10 Ways To Create A Small And Simple Website Without The
Overkill</a></span>
              <span class="link-callout-excerpt">WordPress can be an overkill.
As these other excellent services prove, WordPress is not the be all and end
all of website creation. If you want simpler solutions, there's a variety to
pick from.</span>
```

What is data? Running through some examples…



CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

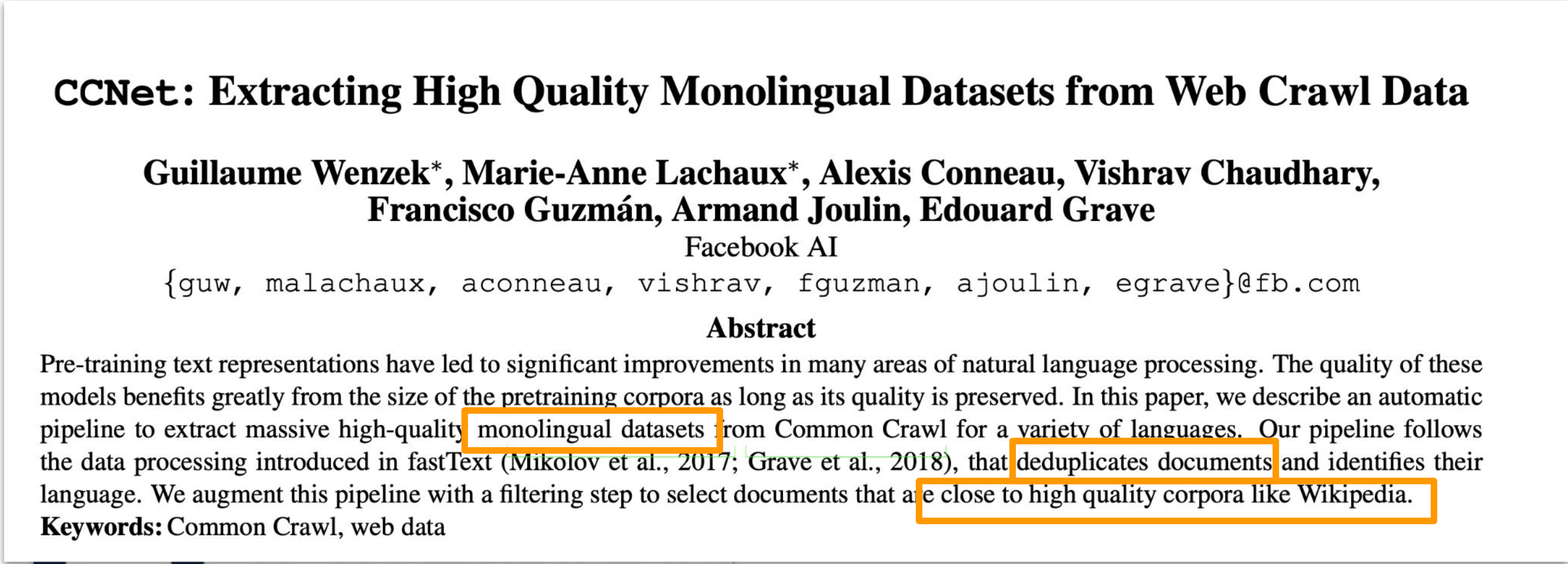
Guillaume Wenzek*, Marie-Anne Lachaux*, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, Edouard Grave
Facebook AI
{guw, malachaux, aconneau, vishrav, fguzman, ajoulin, egrave}@fb.com

Abstract

Pre-training text representations have led to significant improvements in many areas of natural language processing. The quality of these models benefits greatly from the size of the pretraining corpora as long as its quality is preserved. In this paper, we describe an automatic pipeline to extract massive high-quality monolingual datasets from Common Crawl for a variety of languages. Our pipeline follows the data processing introduced in fastText (Mikolov et al., 2017; Grave et al., 2018), that deduplicates documents and identifies their language. We augment this pipeline with a filtering step to select documents that are close to high quality corpora like Wikipedia.

Keywords: Common Crawl, web data

Emerging themes: Language identification, content filtering



CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

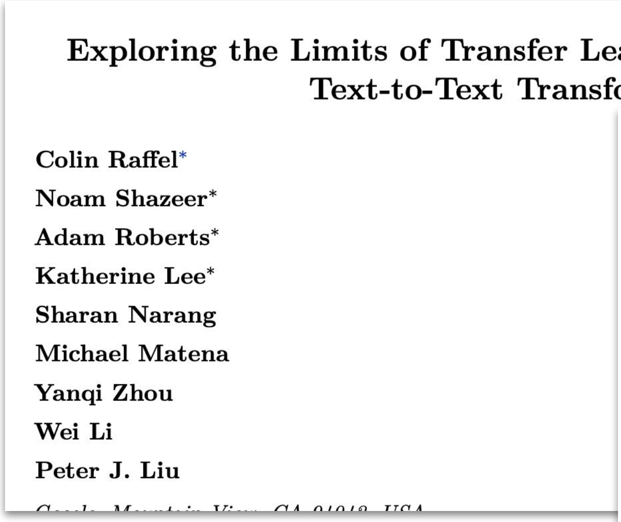
Guillaume Wenzek*, Marie-Anne Lachaux*, Alexis Conneau, Vishrav Chaudhary,
Francisco Guzmán, Armand Joulin, Edouard Grave
Facebook AI
{guw, malachaux, aconneau, vishrav, fguzman, ajoulin, egrave}@fb.com

Abstract

Pre-training text representations have led to significant improvements in many areas of natural language processing. The quality of these models benefits greatly from the size of the pretraining corpora as long as its quality is preserved. In this paper, we describe an automatic pipeline to extract massive high-quality monolingual datasets from Common Crawl for a variety of languages. Our pipeline follows the data processing introduced in fastText (Mikolov et al., 2017; Grave et al., 2018), that deduplicates documents and identifies their language. We augment this pipeline with a filtering step to select documents that are close to high quality corpora like Wikipedia.

Keywords: Common Crawl, web data

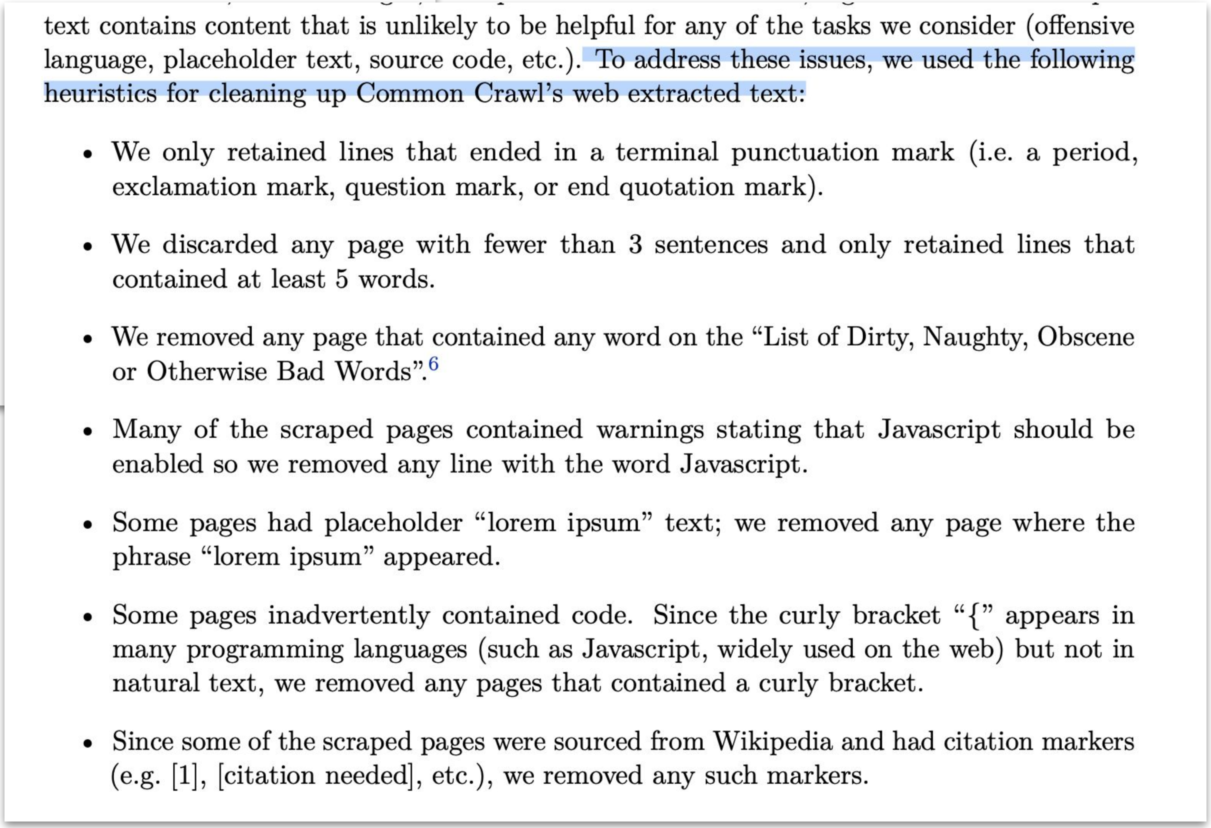
What is data? Running through some examples…



Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*
Noam Shazeer*
Adam Roberts*
Katherine Lee*
Sharan Narang
Michael Matena
Yanqi Zhou
Wei Li
Peter J. Liu

Google, Mountain View, CA 94043, USA

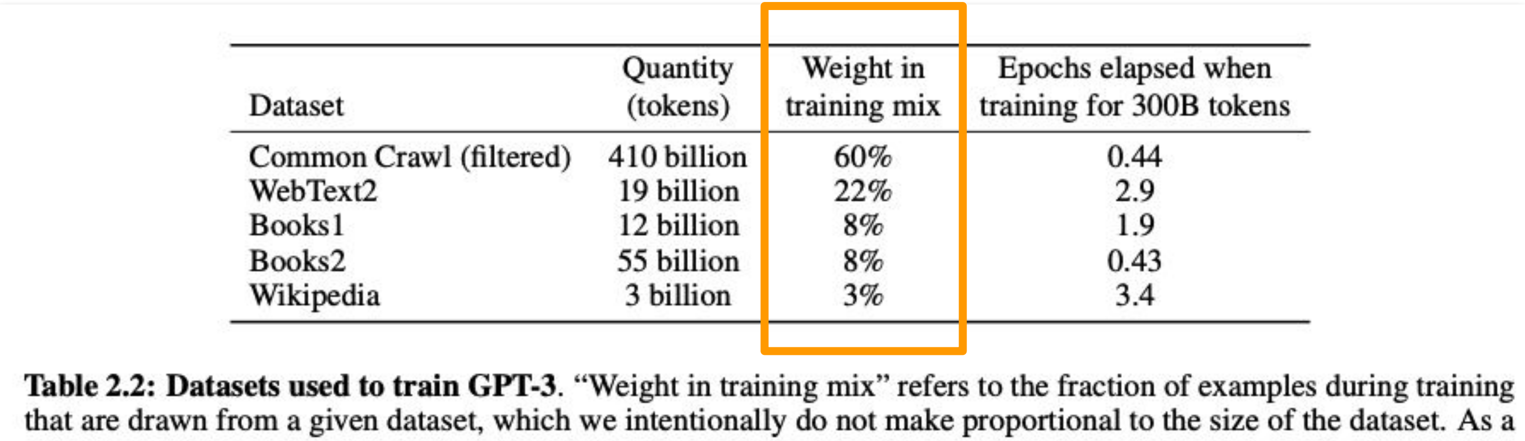


text contains content that is unlikely to be helpful for any of the tasks we consider (offensive language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl's web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- Since some of the scraped pages were sourced from Wikipedia and had citation markers (e.g. [1], [citation needed], etc.), we removed any such markers.

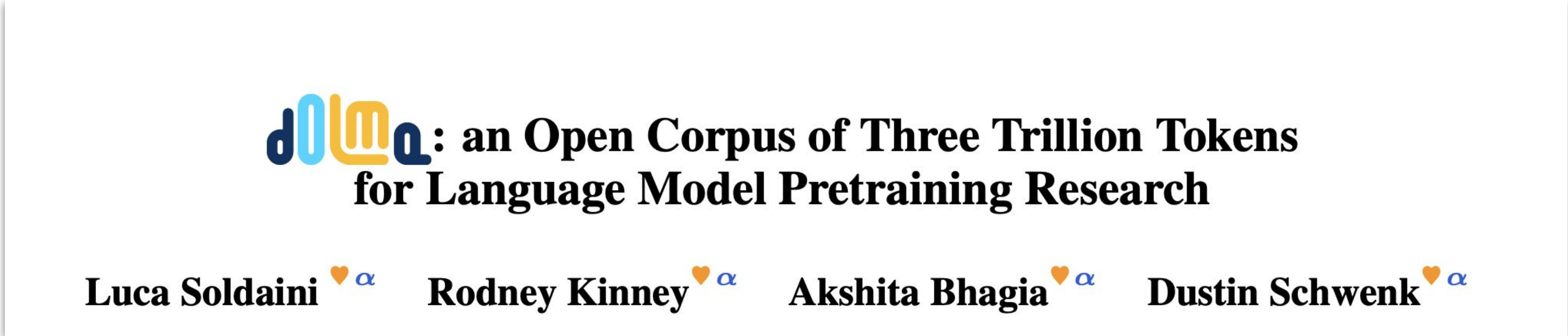
Manual heuristics

Fast forward today: Partial information on data engineering



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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a



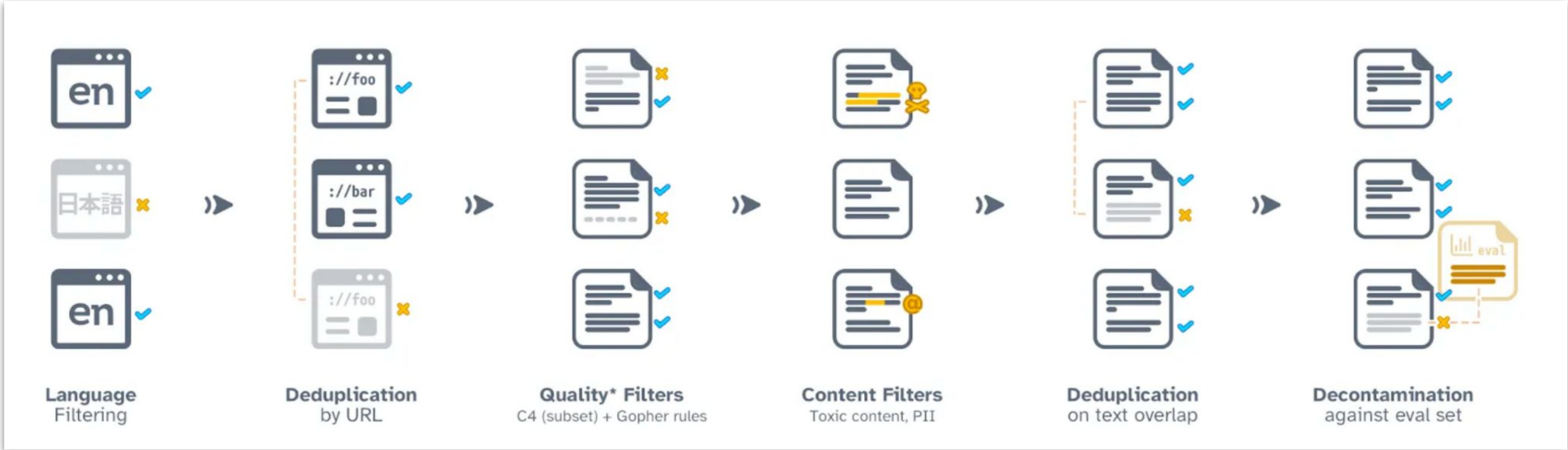
dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research

Luca Soldaini[♥, α] **Rodney Kinney**[♥, α] **Akshita Bhagia**[♥, α] **Dustin Schwenk**[♥, α]

Outline

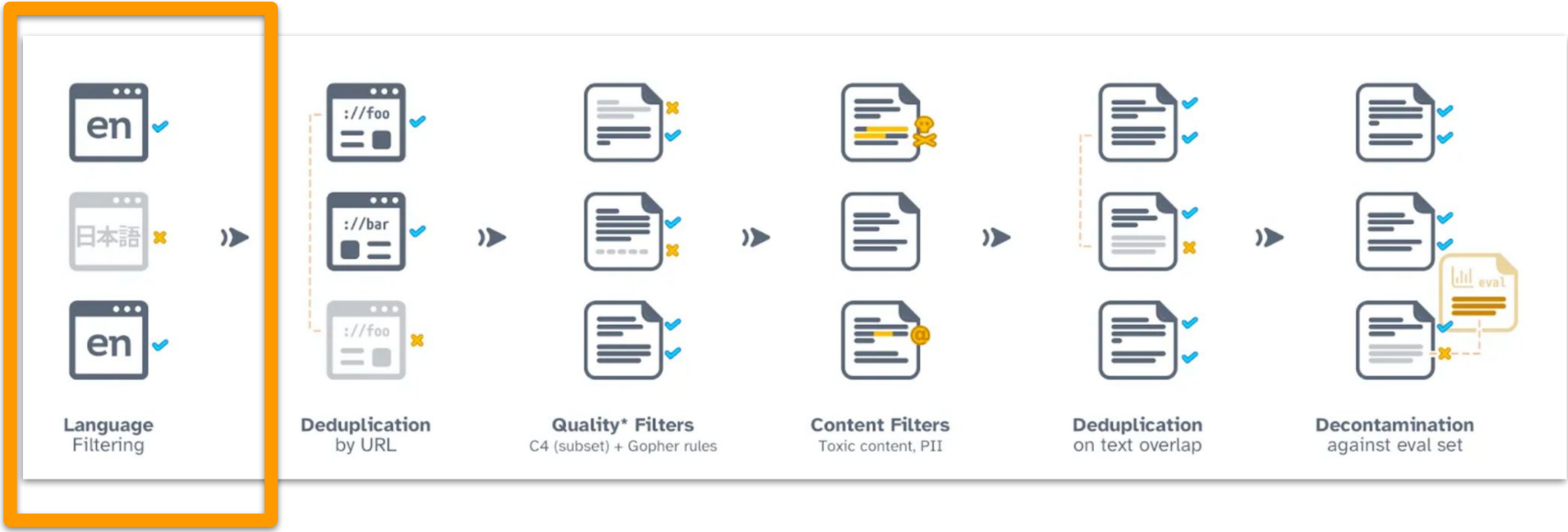
- Why talk about data?
- What is data?
- **How to process data?**
  - Colab!
- What impact does data processing have on downstream applications?
- Hot-take discussion (20 minutes)

Data processing pipeline



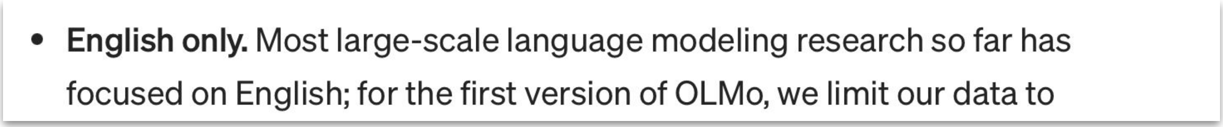
Source: AI2 Dolma

Data processing pipeline



Source: AI2 Dolma

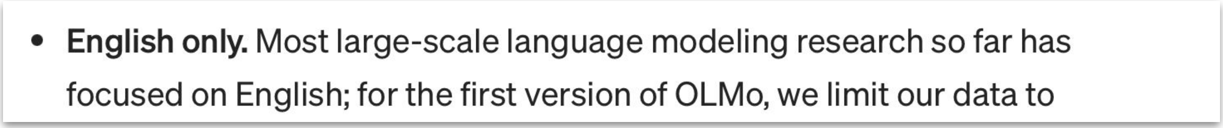
Language identification: Find text in English



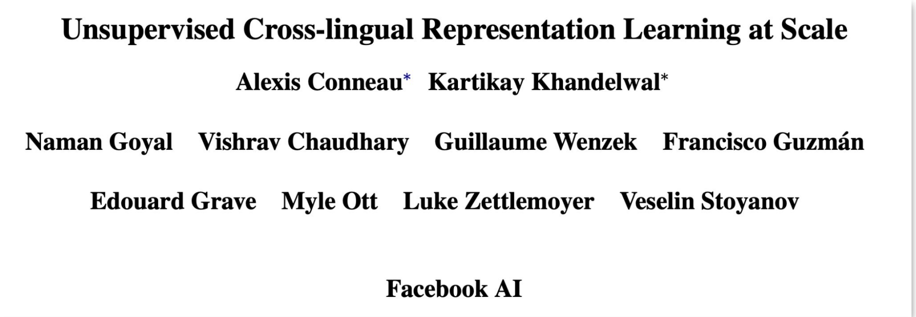
- **English only.** Most large-scale language modeling research so far has focused on English; for the first version of OLMo, we limit our data to

Source: AI2 Dolma

Why monolingual, and not multilingual?



- **English only.** Most large-scale language modeling research so far has focused on English; for the first version of OLMo, we limit our data to



Unsupervised Cross-lingual Representation Learning at Scale

Alexis Conneau* Kartikay Khandelwal*

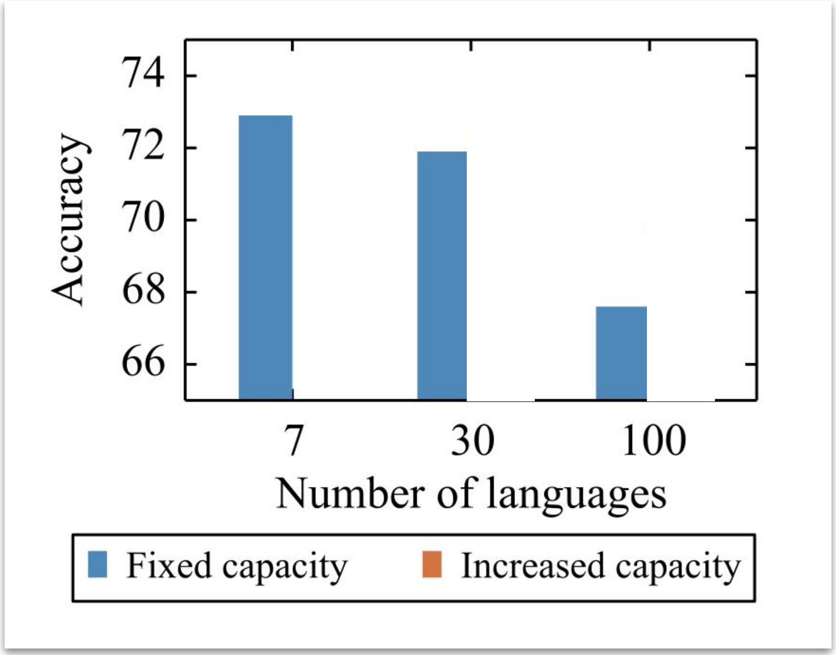
Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán

Edouard Grave Myle Ott Luke Zettlemoyer Veselin Stoyanov

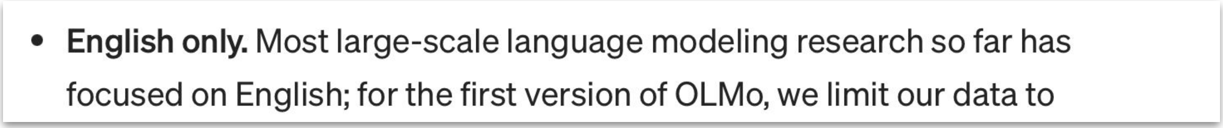
Facebook AI

1. Compute-constrained → less compute per-token over languages.
2. High-quality data curation → challenging per language.

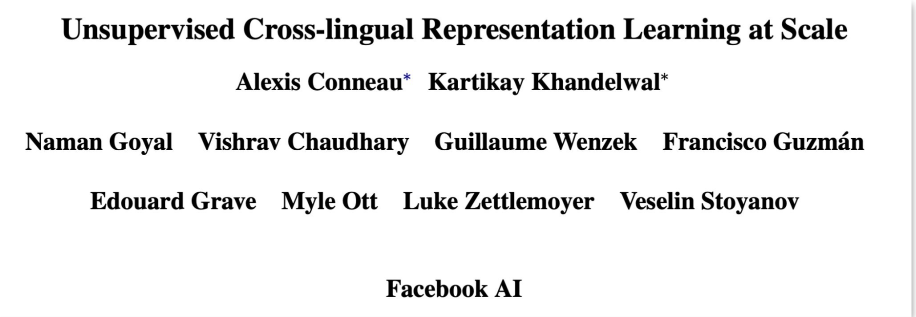
Source: AI2 Dolma



Why monolingual, and not multilingual?



- **English only.** Most large-scale language modeling research so far has focused on English; for the first version of OLMo, we limit our data to



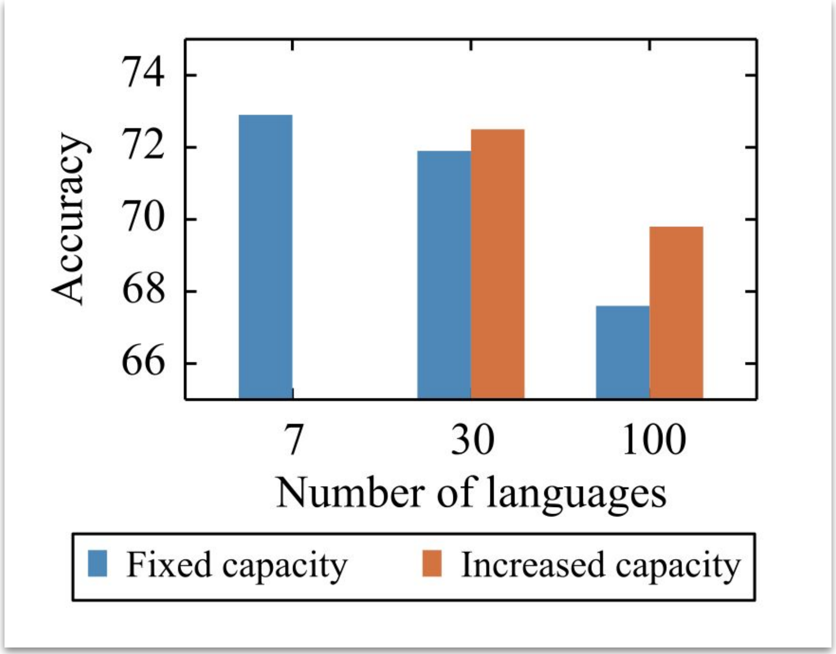
Unsupervised Cross-lingual Representation Learning at Scale

Alexis Conneau* Kartikay Khandelwal*

Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán

Edouard Grave Myle Ott Luke Zettlemoyer Veselin Stoyanov

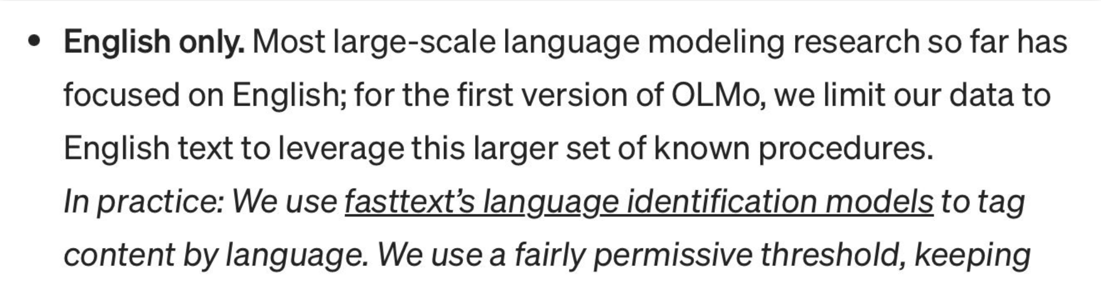
Facebook AI



1. Compute-constrained → less compute per-token over languages.
2. High-quality data curation → challenging per language.

Source: AI2 Dolma

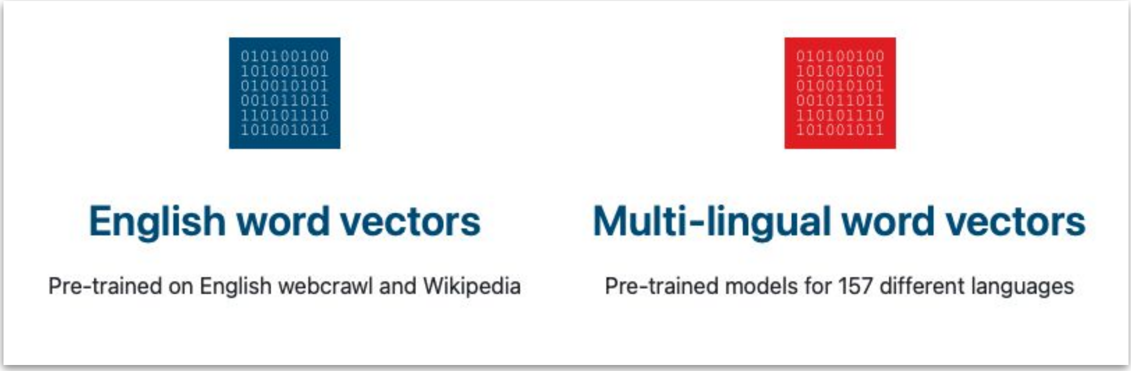
Language identification with fastText



- **English only.** Most large-scale language modeling research so far has focused on English; for the first version of OLMo, we limit our data to English text to leverage this larger set of known procedures.

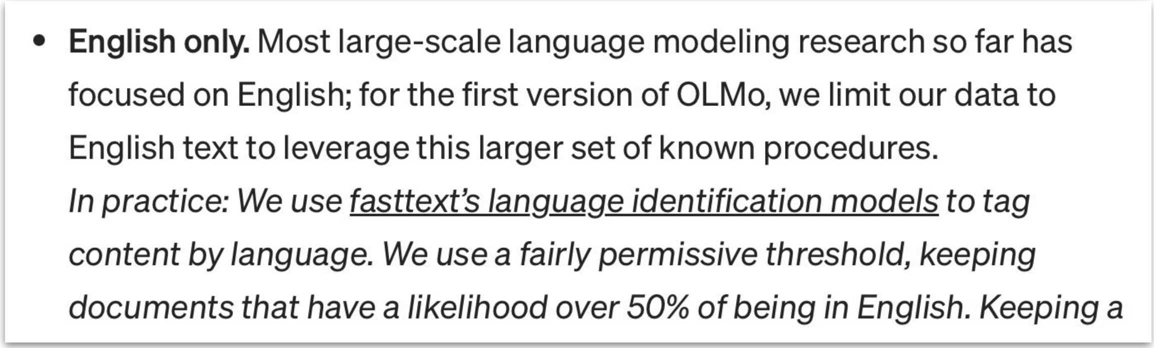
  In practice: We use fasttext's language identification models to tag content by language. We use a fairly permissive threshold, keeping

fastText
Library for efficient text classification and representation learning



Source: AI2 Dolma

Language identification with fastText



- **English only.** Most large-scale language modeling research so far has focused on English; for the first version of OLMo, we limit our data to English text to leverage this larger set of known procedures.
In practice: We use fasttext's language identification models to tag content by language. We use a fairly permissive threshold, keeping documents that have a likelihood over 50% of being in English. Keeping a

Source: AI2 Dolma

We’re going to play a game

Guess the ***language*** and ***score*** for the language!

Colab: [Link]

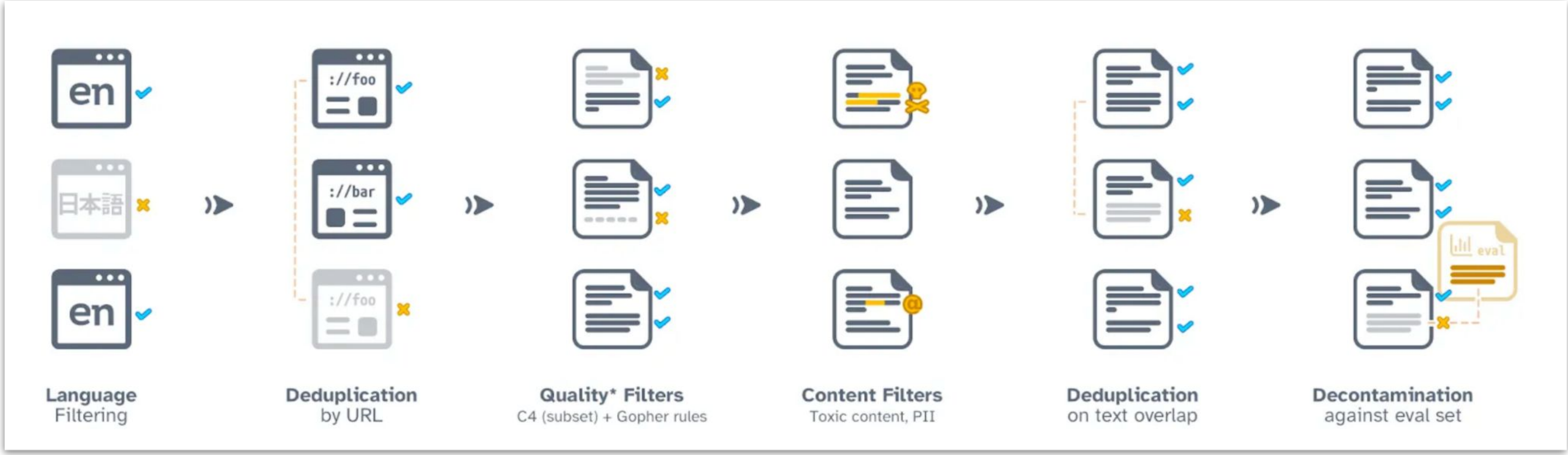
Let’s take a look at the fastText classifier!

How do we feel about this?

From our observations

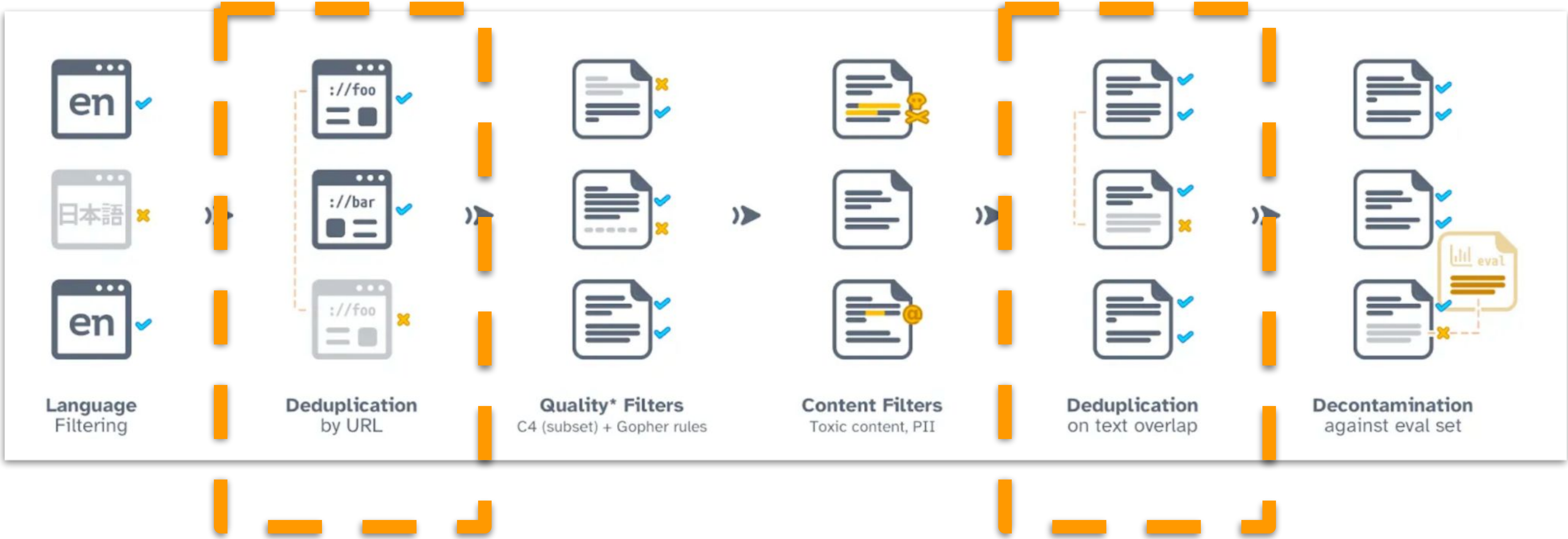
- Scores sometimes feel a bit arbitrary. Tricky when you start to introduce thresholds.
- Tons of edge cases: Difficult on
    - short sequences;
    - English variation e.g., Latex or dialects;
    - code-switching.

Data processing pipeline



Source: AI2 Dolma

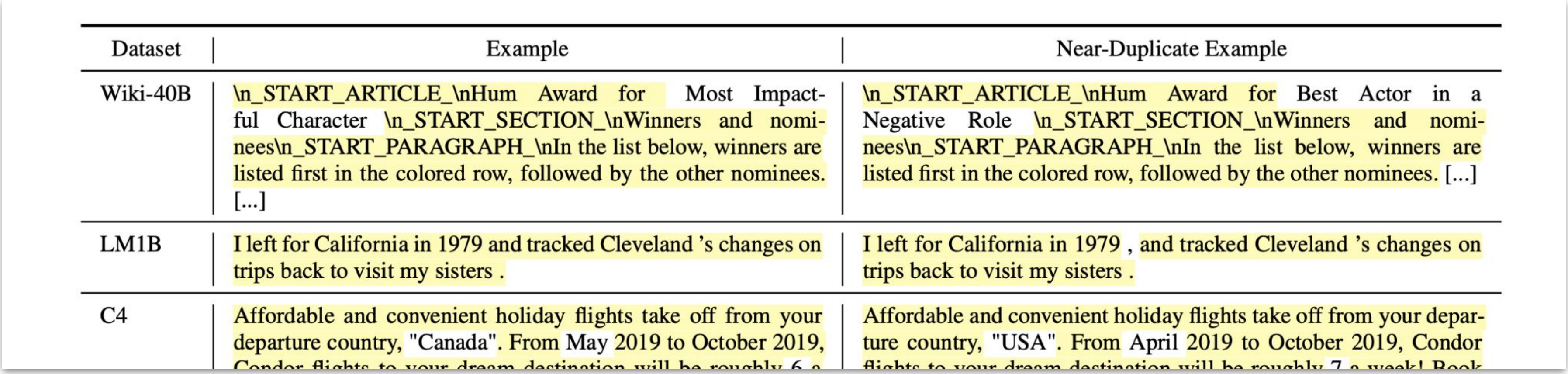
Data processing pipeline



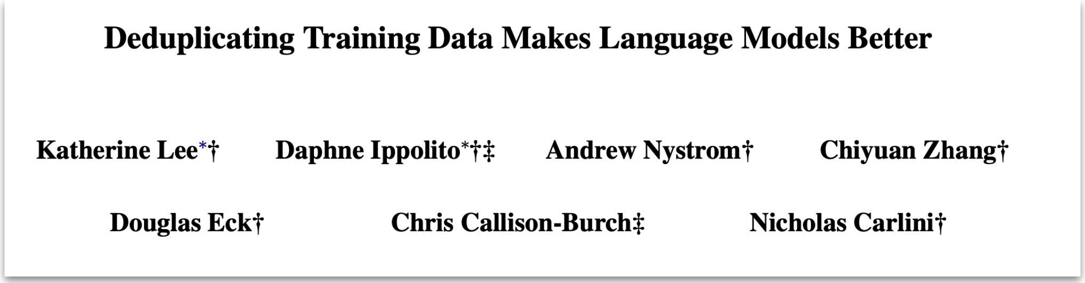
Source: AI2 Dolma

Short comments on deduplication

- Don’t repeat over same *document* (e.g., webpage url). Downstream: Less text memorization.
- Don’t repeat over similar *text content* (e.g., below)



Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland ’s changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland ’s changes on trips back to visit my sisters .
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book

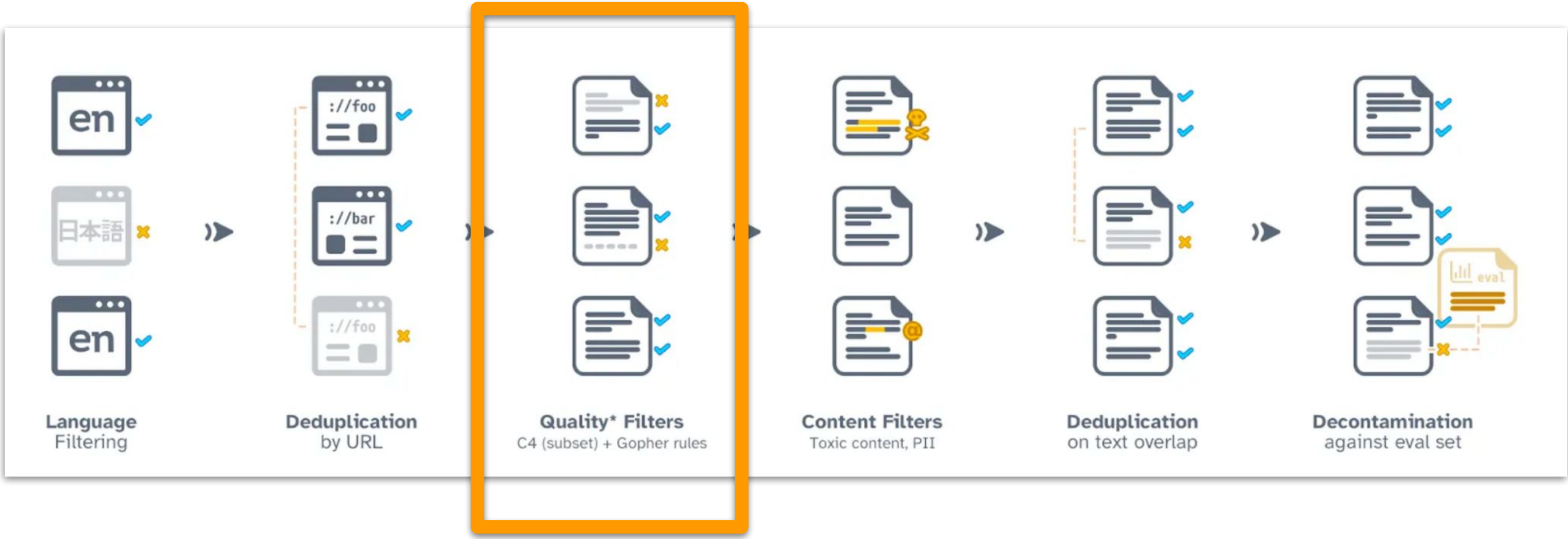


Deduplicating Training Data Makes Language Models Better

Katherine Lee*† Daphne Ippolito*†‡ Andrew Nystrom† Chiyuan Zhang†

Douglas Eck† Chris Callison-Burch‡ Nicholas Carlini†

Data processing pipeline

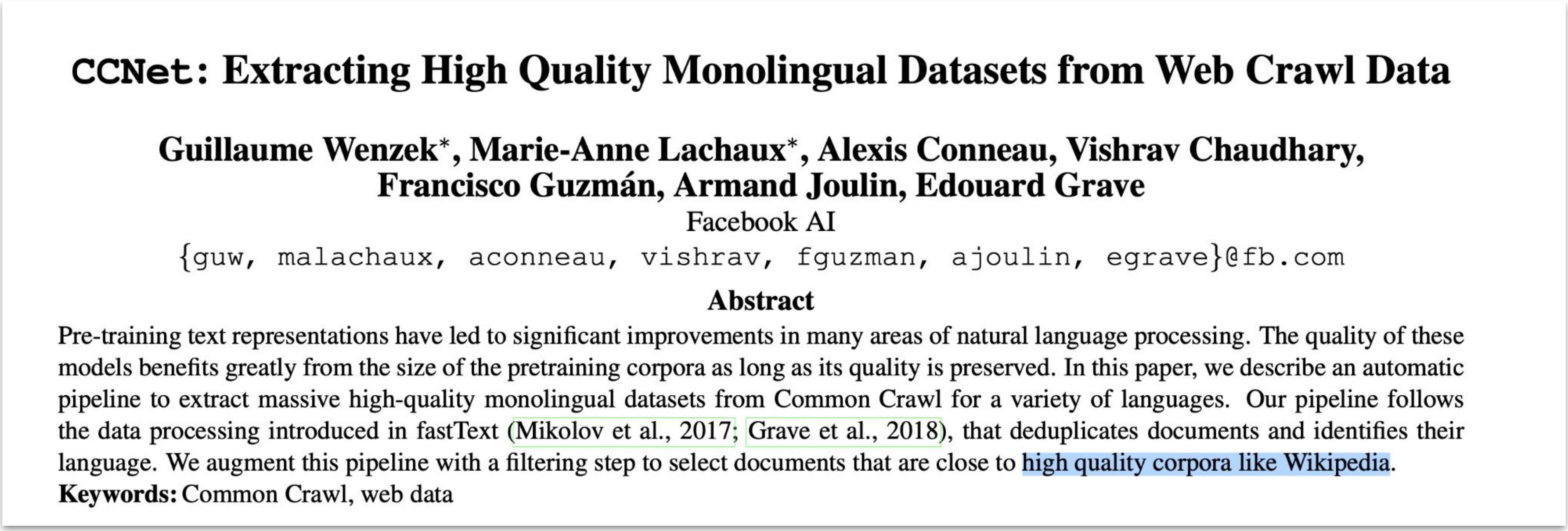


Source: AI2 Dolma

Data Quality

- Lots of web data is not “high-quality”.
- *Definition* of quality varies.
- *Detection* method of quality varies.

Define High-Quality Data = Wikipedia-like



CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

Guillaume Wenzek*, Marie-Anne Lachaux*, Alexis Conneau, Vishrav Chaudhary,
Francisco Guzmán, Armand Joulin, Edouard Grave
Facebook AI
{guw, malachaux, aconneau, vishrav, fguzman, ajoulin, egrave}@fb.com

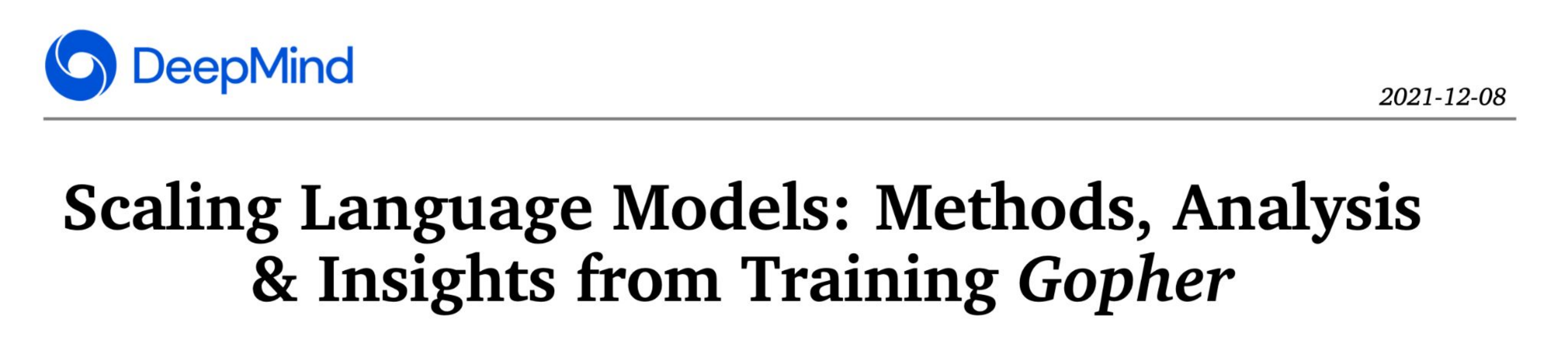
Abstract

Pre-training text representations have led to significant improvements in many areas of natural language processing. The quality of these models benefits greatly from the size of the pretraining corpora as long as its quality is preserved. In this paper, we describe an automatic pipeline to extract massive high-quality monolingual datasets from Common Crawl for a variety of languages. Our pipeline follows the data processing introduced in fastText (Mikolov et al., 2017; Grave et al., 2018), that deduplicates documents and identifies their language. We augment this pipeline with a filtering step to select documents that are close to high quality corpora like Wikipedia.

Keywords: Common Crawl, web data

No, wait!

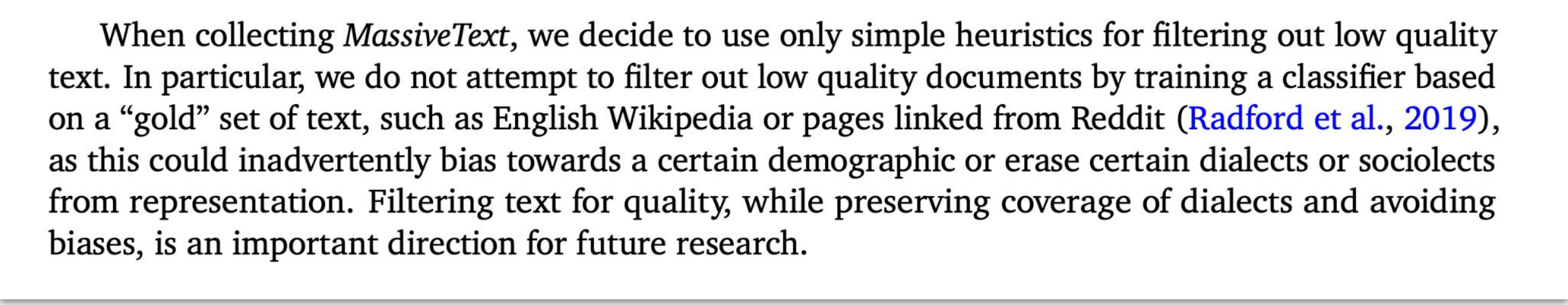
Define High-Quality Data ≠ Wikipedia-like



DeepMind

2021-12-08

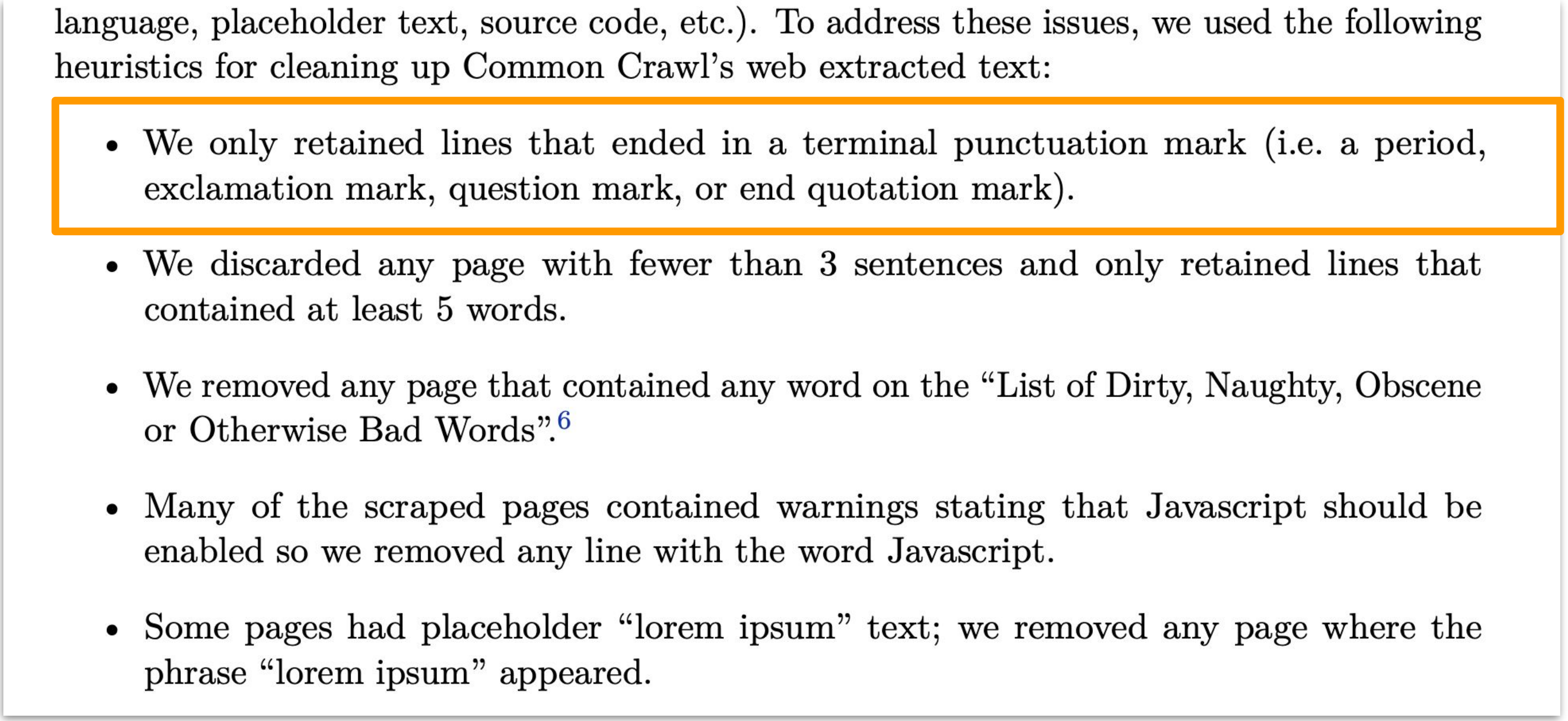
Scaling Language Models: Methods, Analysis & Insights from Training *Gopher*



When collecting *MassiveText*, we decide to use only simple heuristics for filtering out low quality text. In particular, we do not attempt to filter out low quality documents by training a classifier based on a "gold" set of text, such as English Wikipedia or pages linked from Reddit (Radford et al., 2019), as this could inadvertently bias towards a certain demographic or erase certain dialects or sociolects from representation. Filtering text for quality, while preserving coverage of dialects and avoiding biases, is an important direction for future research.

No, wait!

Define High-Quality Data = Intuition & Heuristics



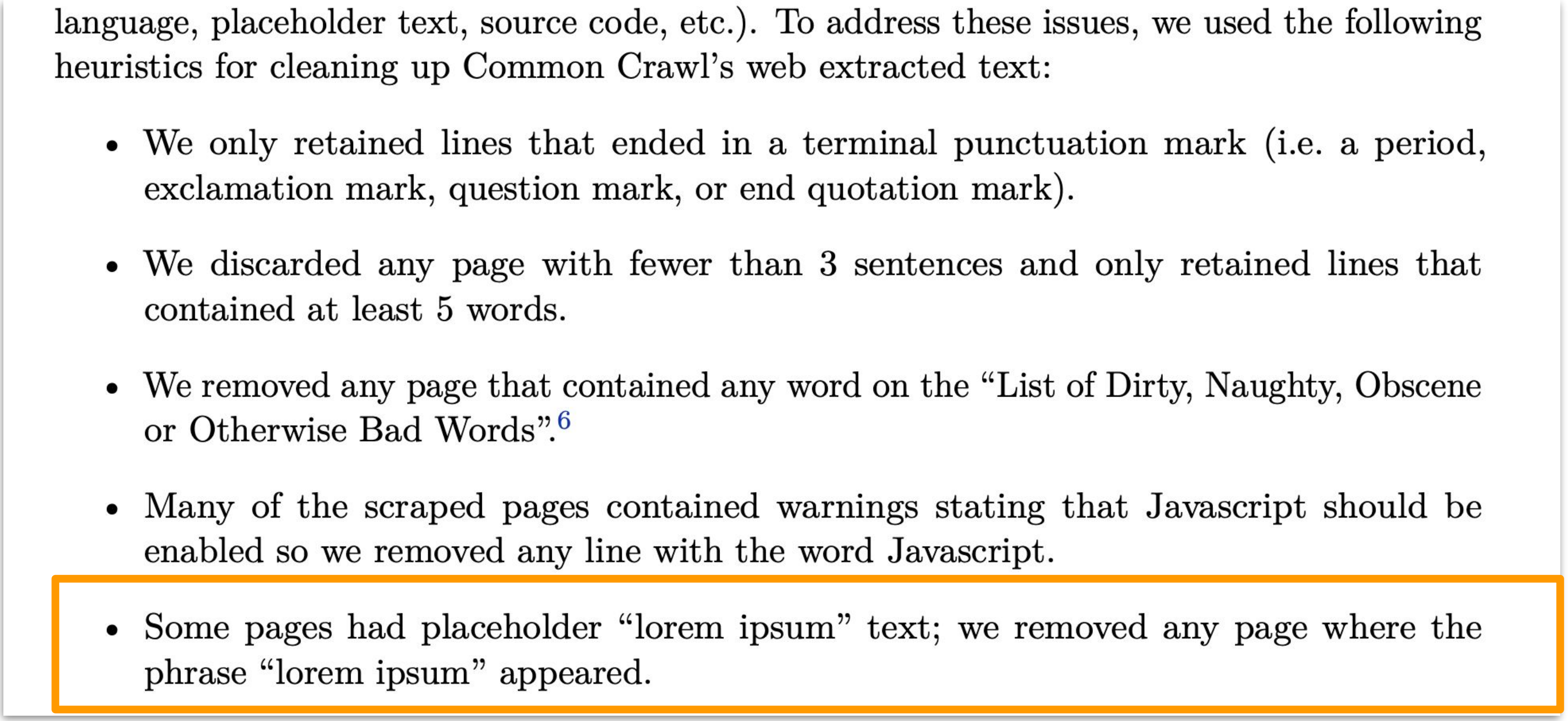
language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl's web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.

C4 dataset

No, wait!

Define High-Quality Data = Intuition & Heuristics



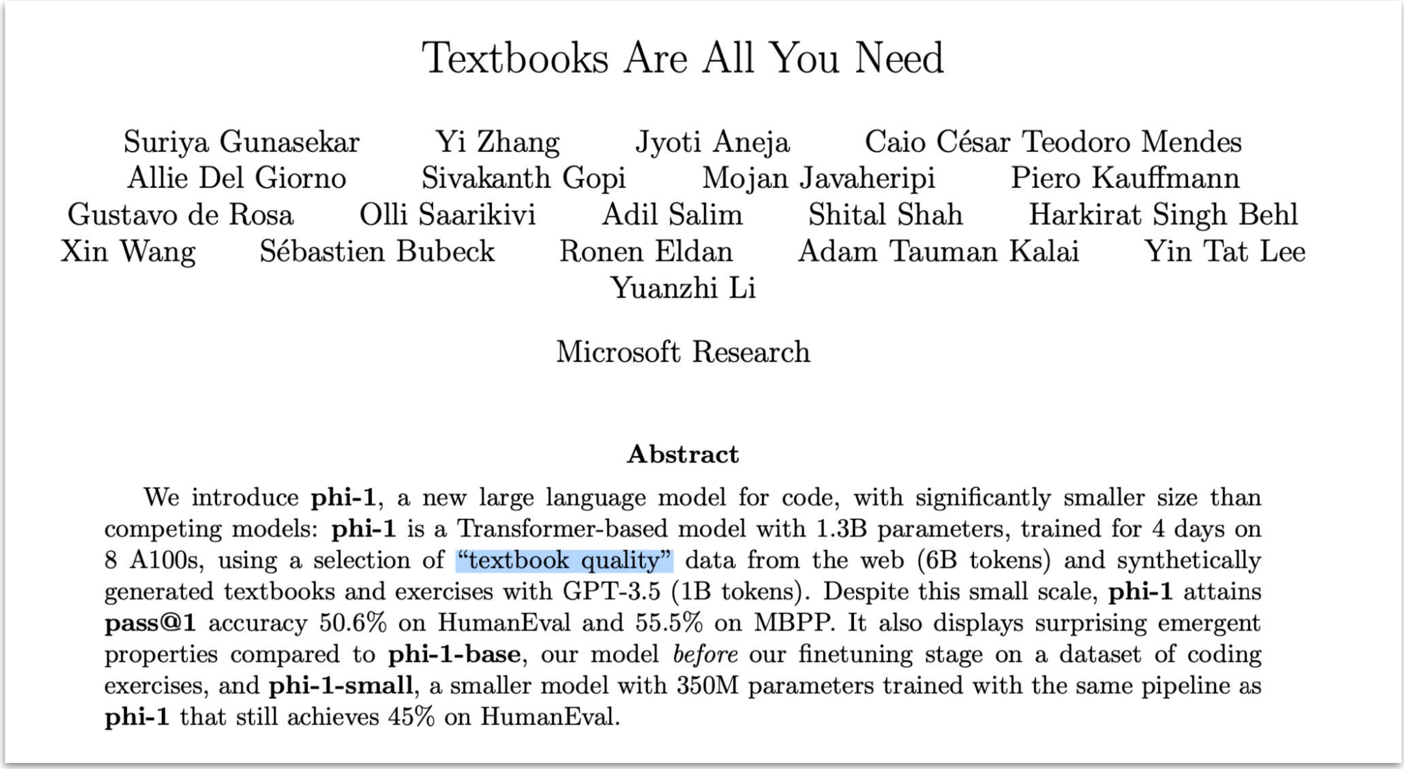
language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl's web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words”.[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder “lorem ipsum” text; we removed any page where the phrase “lorem ipsum” appeared.

C4 dataset

No, wait!

Data Quality = Textbook-like



Textbooks Are All You Need

Suriya Gunasekar Yi Zhang Jyoti Aneja Caio César Teodoro Mendes
Allie Del Giorno Sivakanth Gopi Mojan Javaheripi Piero Kauffmann
Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Harkirat Singh Behl
Xin Wang Sébastien Bubeck Ronen Eldan Adam Tauman Kalai Yin Tat Lee
Yuanzhi Li

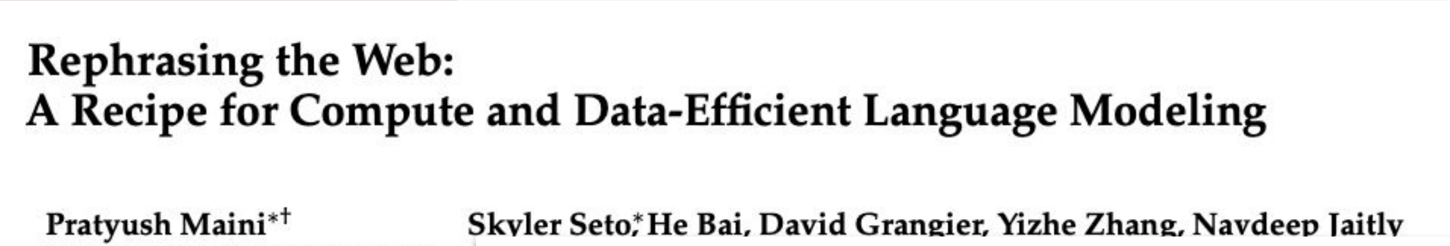
Microsoft Research

Abstract

We introduce **phi-1**, a new large language model for code, with significantly smaller size than competing models: **phi-1** is a Transformer-based model with 1.3B parameters, trained for 4 days on 8 A100s, using a selection of "textbook quality" data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens). Despite this small scale, **phi-1** attains **pass@1** accuracy 50.6% on HumanEval and 55.5% on MBPP. It also displays surprising emergent properties compared to **phi-1-base**, our model *before* our finetuning stage on a dataset of coding exercises, and **phi-1-small**, a smaller model with 350M parameters trained with the same pipeline as **phi-1** that still achieves 45% on HumanEval.

No, wait!

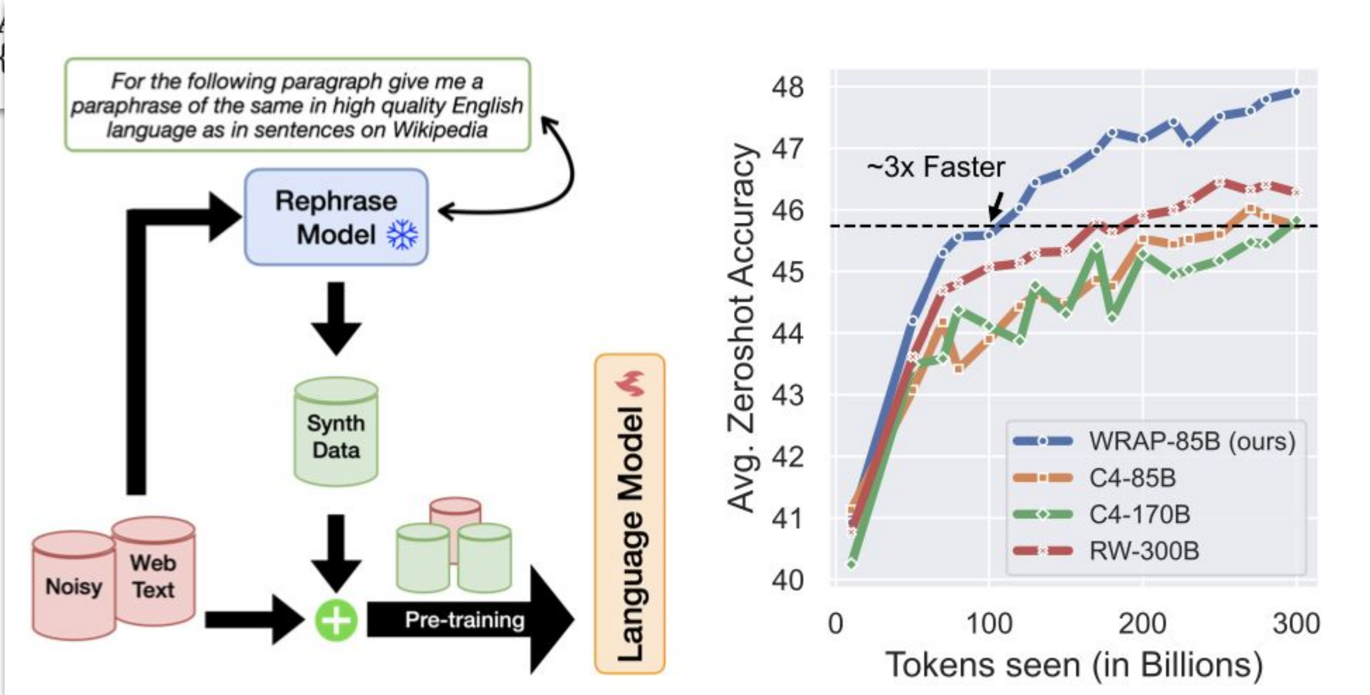
Synthetic approaches



Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling

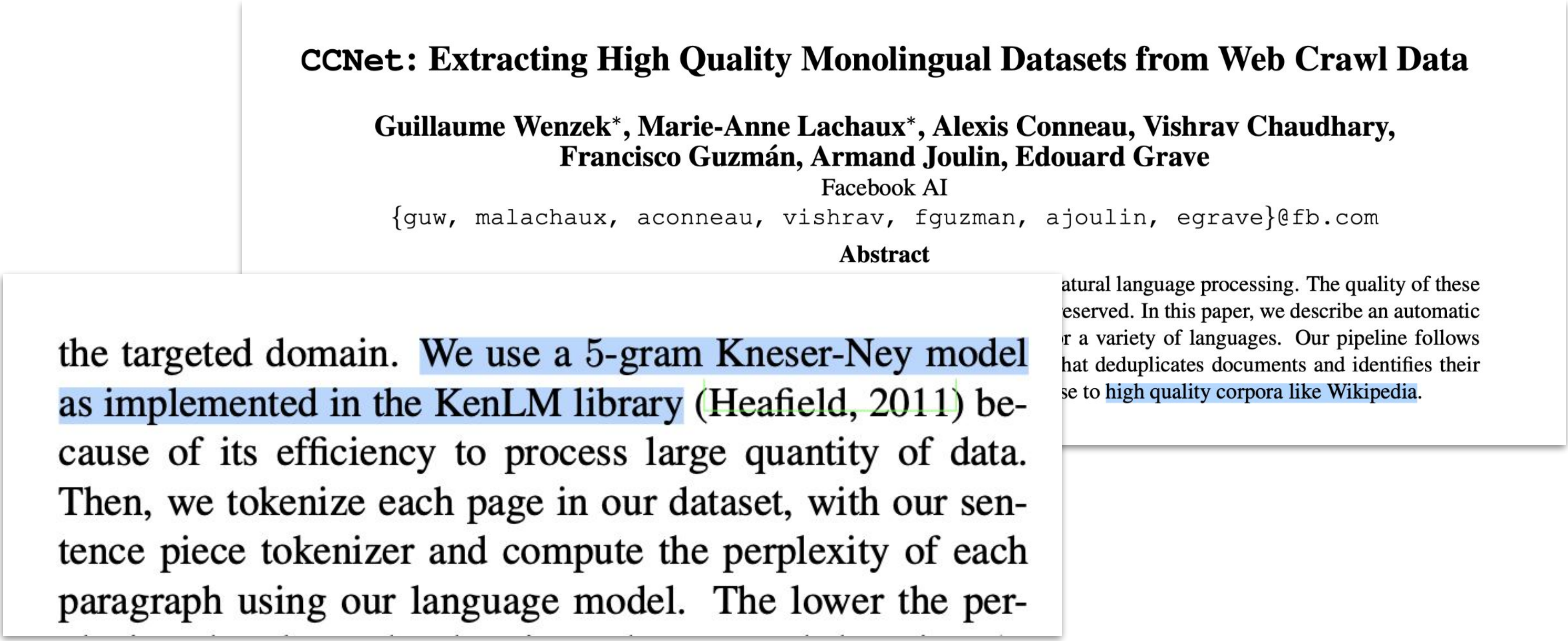
Pratyush Maini[*†]
Carnegie Mellon Univeristy
pratyushmaini@cmu.edu

Skyler Seto,[*] He Bai, David Grangier, Yizhe Zhang, Navdeep Jaitly
A
{



What about *detecting* for high-quality data at scale?

Detecting high-quality data (e.g., Wikipedia)



CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

Guillaume Wenzek*, Marie-Anne Lachaux*, Alexis Conneau, Vishrav Chaudhary,
Francisco Guzmán, Armand Joulin, Edouard Grave
Facebook AI
{guw, malachaux, aconneau, vishrav, fguzman, ajoulin, egrave}@fb.com

Abstract

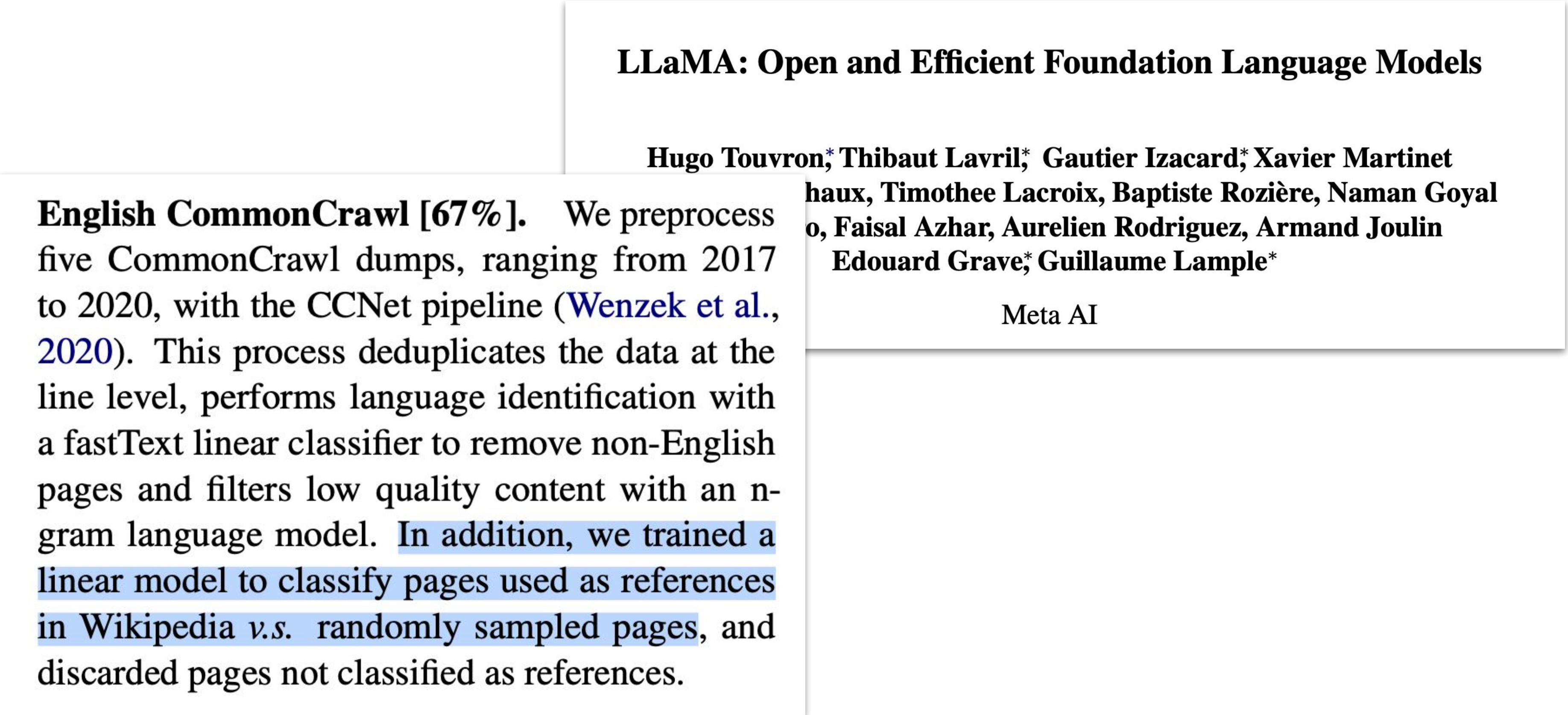
atural language processing. The quality of these
eserved. In this paper, we describe an automatic
r a variety of languages. Our pipeline follows
hat deduplicates documents and identifies their
se to high quality corpora like Wikipedia.

the targeted domain. We use a 5-gram Kneser-Ney model as implemented in the KenLM library (Heafield, 2011) because of its efficiency to process large quantity of data. Then, we tokenize each page in our dataset, with our sentence piece tokenizer and compute the perplexity of each paragraph using our language model. The lower the per-

Detecting high-quality data (e.g., Wikipedia)

Example from the same Llama/Meta team

Detecting high-quality data (e.g., Wikipedia)



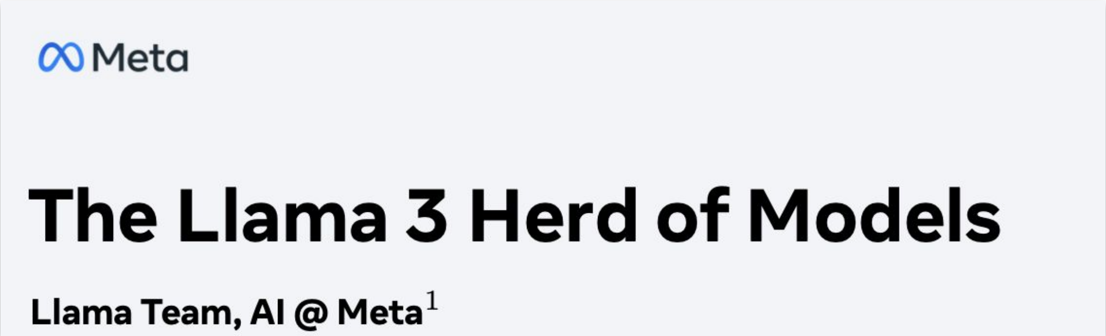
LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron,* Thibaut Lavril,* Gautier Izacard,* Xavier Martinet
haux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
o, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave,* Guillaume Lample*

Meta AI

English CommonCrawl [67%]. We preprocess five CommonCrawl dumps, ranging from 2017 to 2020, with the CCNet pipeline (Wenzek et al., 2020). This process deduplicates the data at the line level, performs language identification with a fastText linear classifier to remove non-English pages and filters low quality content with an n-gram language model. In addition, we trained a linear model to classify pages used as references in Wikipedia *v.s.* randomly sampled pages, and discarded pages not classified as references.

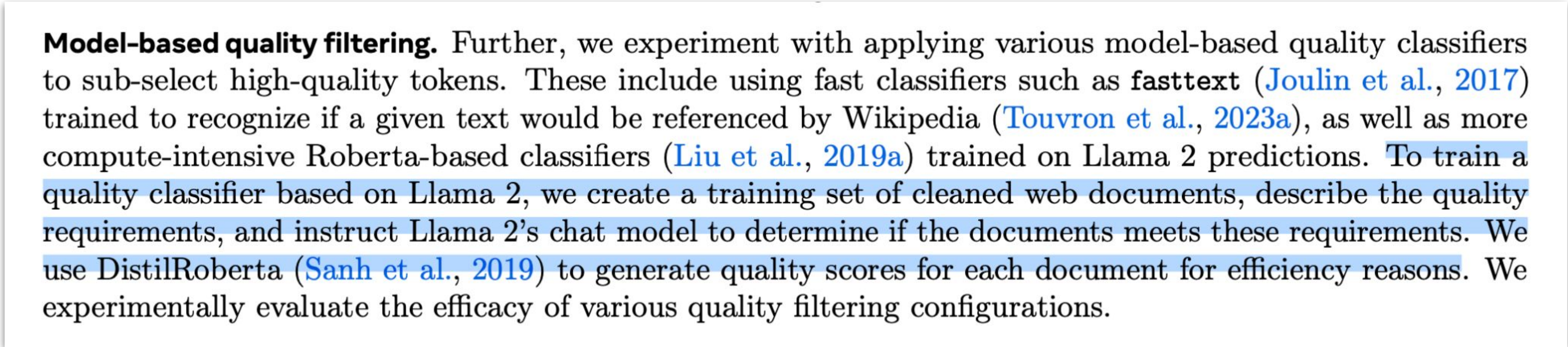
Detecting high-quality data (e.g., Wikipedia)



Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]



Model-based quality filtering. Further, we experiment with applying various model-based quality classifiers to sub-select high-quality tokens. These include using fast classifiers such as **fasttext** (Joulin et al., 2017) trained to recognize if a given text would be referenced by Wikipedia (Touvron et al., 2023a), as well as more compute-intensive Roberta-based classifiers (Liu et al., 2019a) trained on Llama 2 predictions. To train a quality classifier based on Llama 2, we create a training set of cleaned web documents, describe the quality requirements, and instruct Llama 2's chat model to determine if the documents meets these requirements. We use DistilRoberta (Sanh et al., 2019) to generate quality scores for each document for efficiency reasons. We experimentally evaluate the efficacy of various quality filtering configurations.

Defining and detecting high-quality is HARD.

Let’s play another game: Guess the ***quality-score***!

Assume high-quality = Wikipedia like.

Colab: [Link]

BREAK 5 mins

Stretch.

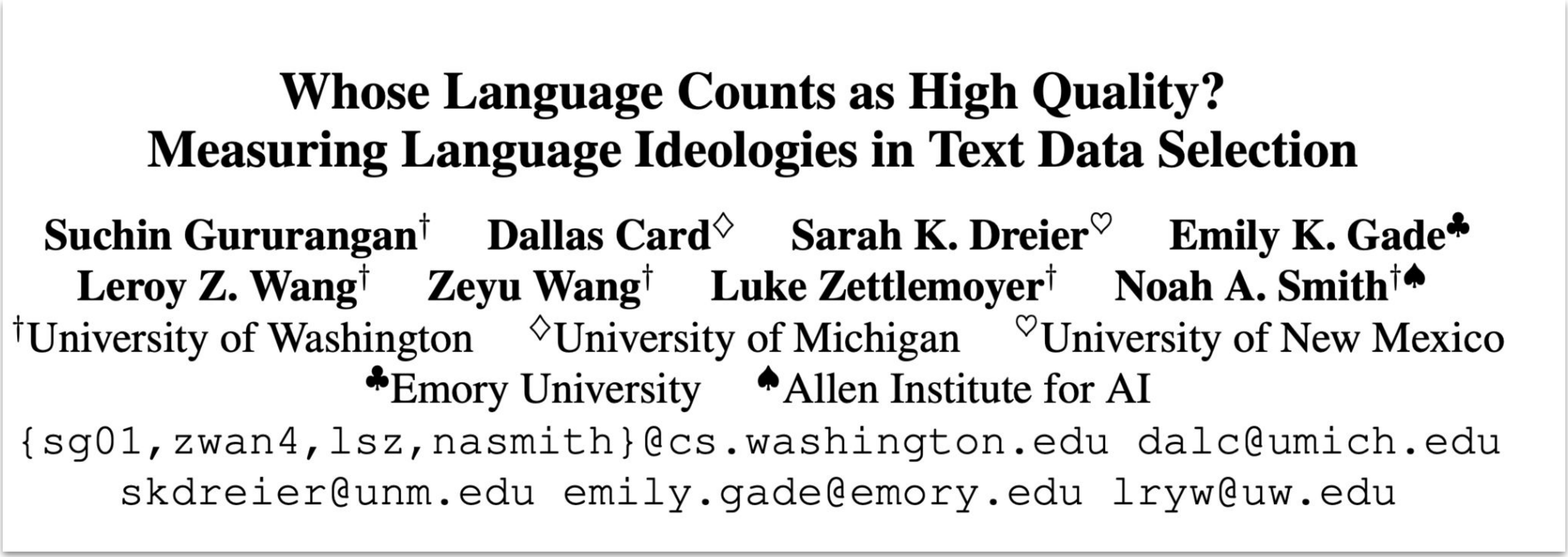
Or play with the Colab :-)

http://goto.stanford.edu/rewang-data

Outline

- Why talk about data?
- What is data?
- How to process data?
  - Colab!
- **What impact does data processing have on downstream applications?**
- Hot-take discussion (20 minutes)

Implications of data choices? Case study on quality filter.



Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection

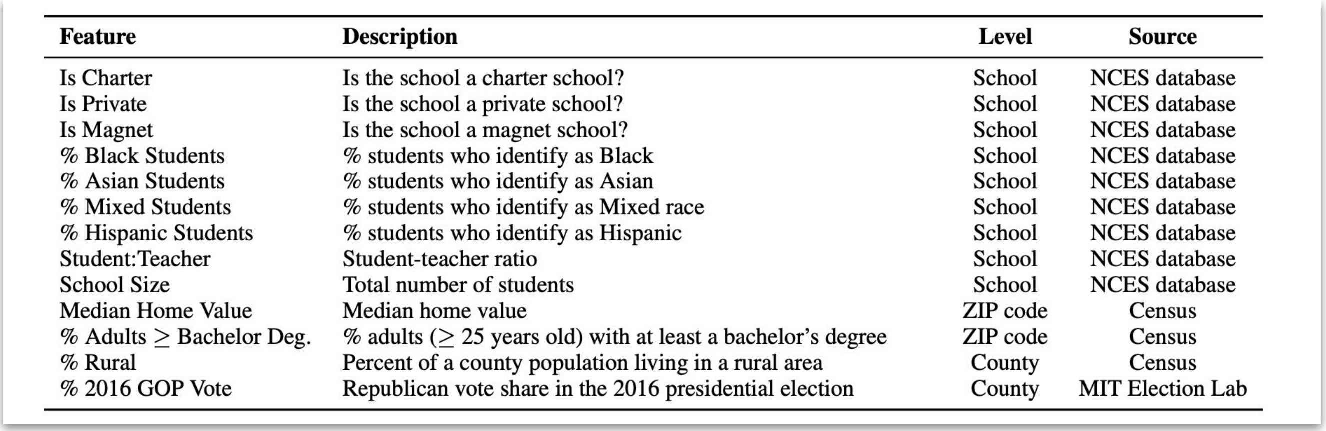
Suchin Gururangan[†] Dallas Card[◇] Sarah K. Dreier[♡] Emily K. Gade[♣]
Leroy Z. Wang[†] Zeyu Wang[†] Luke Zettlemoyer[†] Noah A. Smith[†♠]
[†]University of Washington [◇]University of Michigan [♡]University of New Mexico
[♣]Emory University [♠]Allen Institute for AI
{sg01,zwan4,lsz,nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

Implications of quality filter

Replicating the GPT-3 quality filter

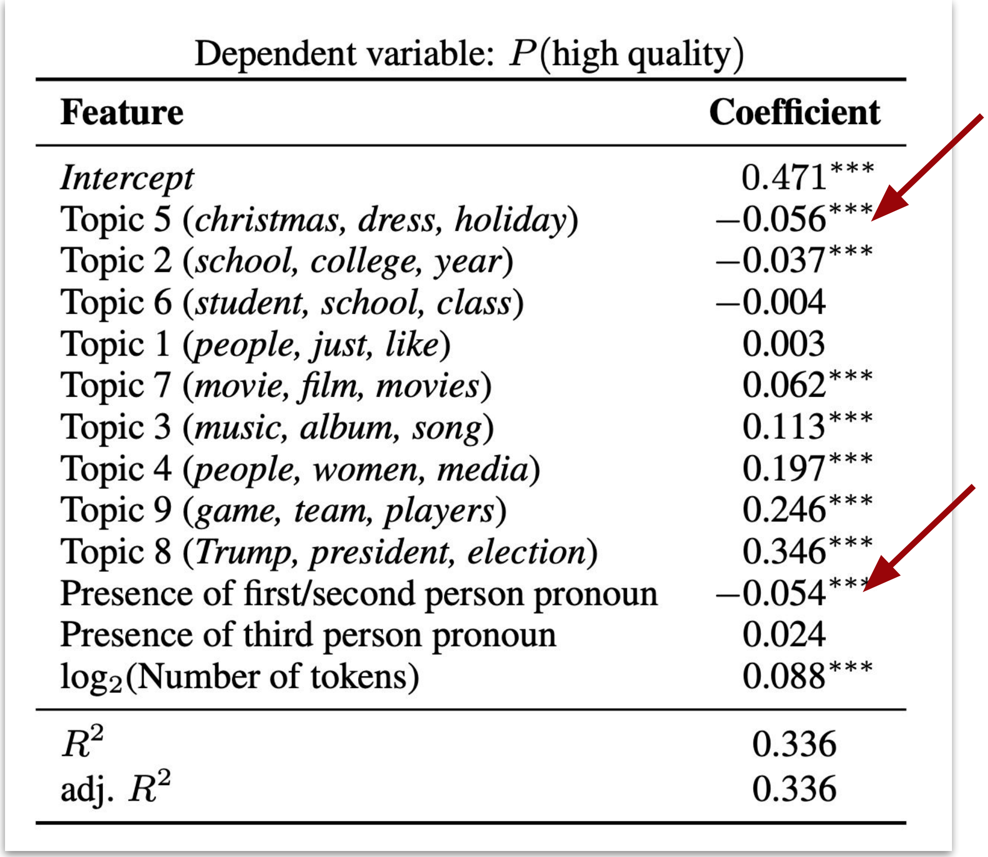
- Books3, Wikipedia, OpenWebText vs. random sample of Common Crawl
- 90.4% F1 score

Applied to U.S. high school newspapers linked with human factors.



Feature	Description	Level	Source
Is Charter	Is the school a charter school?	School	NCES database
Is Private	Is the school a private school?	School	NCES database
Is Magnet	Is the school a magnet school?	School	NCES database
% Black Students	% students who identify as Black	School	NCES database
% Asian Students	% students who identify as Asian	School	NCES database
% Mixed Students	% students who identify as Mixed race	School	NCES database
% Hispanic Students	% students who identify as Hispanic	School	NCES database
Student:Teacher	Student-teacher ratio	School	NCES database
School Size	Total number of students	School	NCES database
Median Home Value	Median home value	ZIP code	Census
% Adults $\geq$ Bachelor Deg.	% adults ($\geq$ 25 years old) with at least a bachelor's degree	ZIP code	Census
% Rural	Percent of a county population living in a rural area	County	Census
% 2016 GOP Vote	Republican vote share in the 2016 presidential election	County	MIT Election Lab

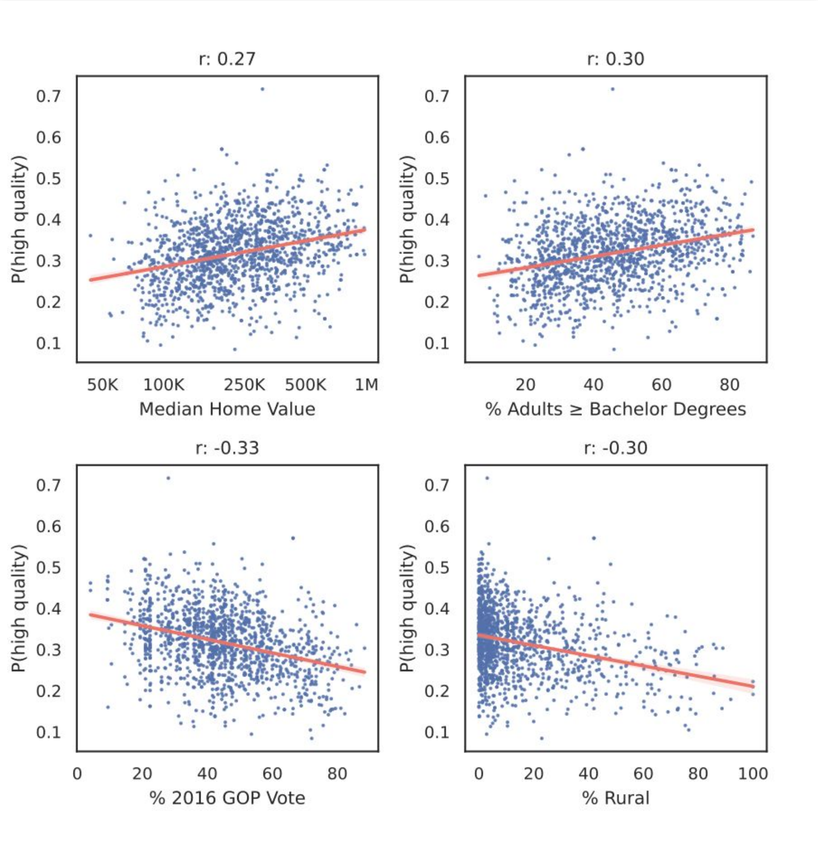
What topics are considered “high quality”?



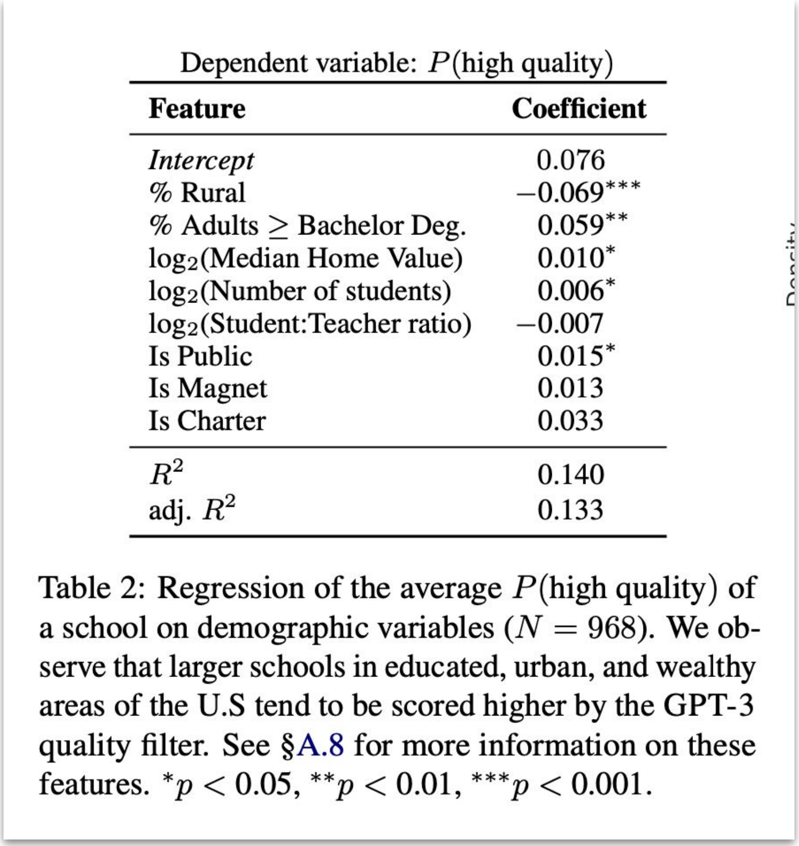
Dependent variable: P(high quality)

Feature	**Coefficient**
Intercept	0.471^{***}
Topic 5 (*christmas, dress, holiday*)	-0.056^{***}
Topic 2 (*school, college, year*)	-0.037^{***}
Topic 6 (*student, school, class*)	-0.004
Topic 1 (*people, just, like*)	0.003
Topic 7 (*movie, film, movies*)	0.062^{***}
Topic 3 (*music, album, song*)	0.113^{***}
Topic 4 (*people, women, media*)	0.197^{***}
Topic 9 (*game, team, players*)	0.246^{***}
Topic 8 (*Trump, president, election*)	0.346^{***}
Presence of first/second person pronoun	-0.054^{***}
Presence of third person pronoun	0.024
$\log_2$(Number of tokens)	0.088^{***}
R^2	0.336
adj. R^2	0.336

Who is considered “high-quality”?



Who is considered “high-quality”?



Dependent variable: P(high quality)

Feature	**Coefficient**
Intercept	0.076
% Rural	-0.069^{***}
% Adults $\geq$ Bachelor Deg.	0.059^{**}
$\log_2$(Median Home Value)	0.010^{*}
$\log_2$(Number of students)	0.006^{*}
$\log_2$(Student:Teacher ratio)	-0.007
Is Public	0.015^{*}
Is Magnet	0.013
Is Charter	0.033
R^2	0.140
adj. R^2	0.133

Table 2: Regression of the average P(high quality) of a school on demographic variables ($N = 968$). We observe that larger schools in educated, urban, and wealthy areas of the U.S tend to be scored higher by the GPT-3 quality filter. See §A.8 for more information on these features. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

But…is it because bad text quality?

But…is it because bad text quality?

No!

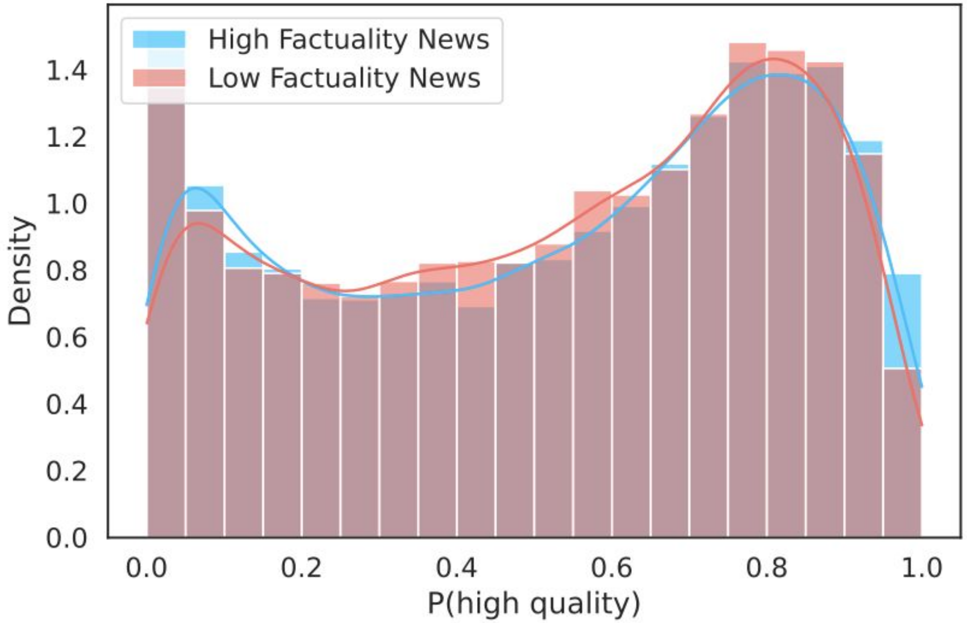


Figure 2: There is no difference in quality scores between articles written by news sources of high and low factual reliability.

Summary

- Data collection is hard; there are a lot of *ad-hoc choices* and unreliable methods.
  - e.g., the way we filter data by language or “quality”
- These choices and decisions have real downstream implications on us!

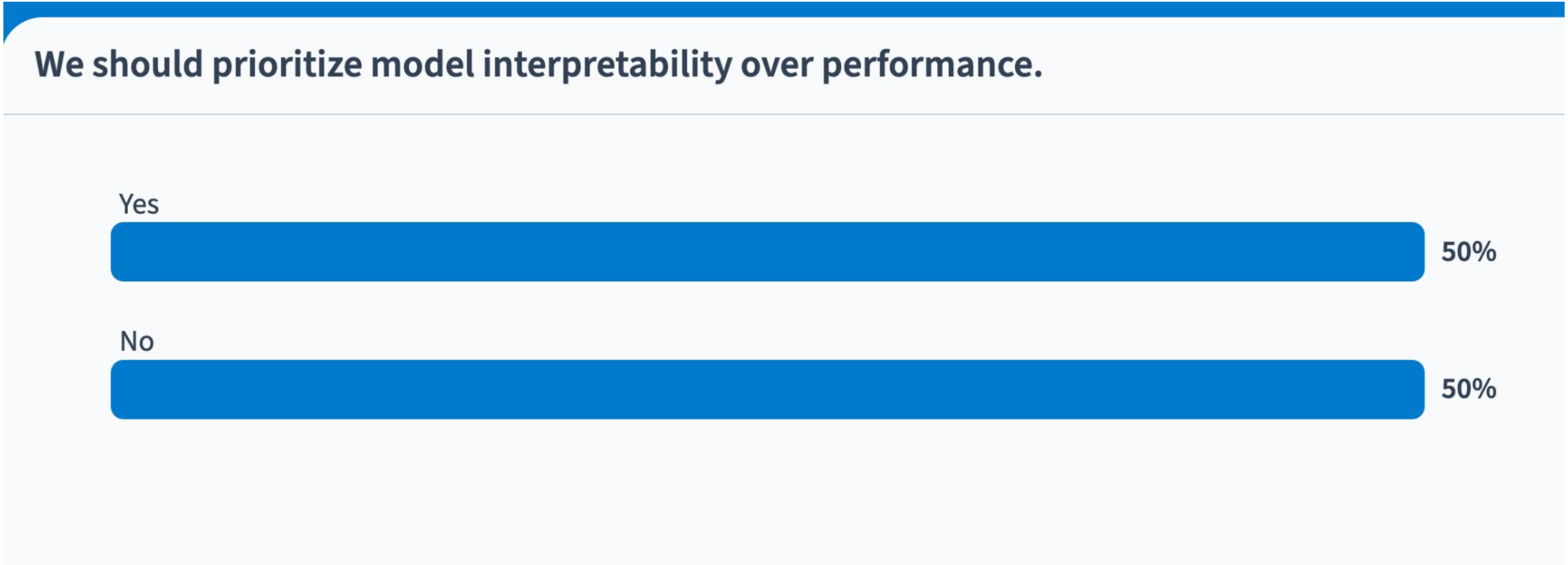
Hot-take

PollEv.com/calebziems988

We should prioritize model interpretability over performance.

02:00

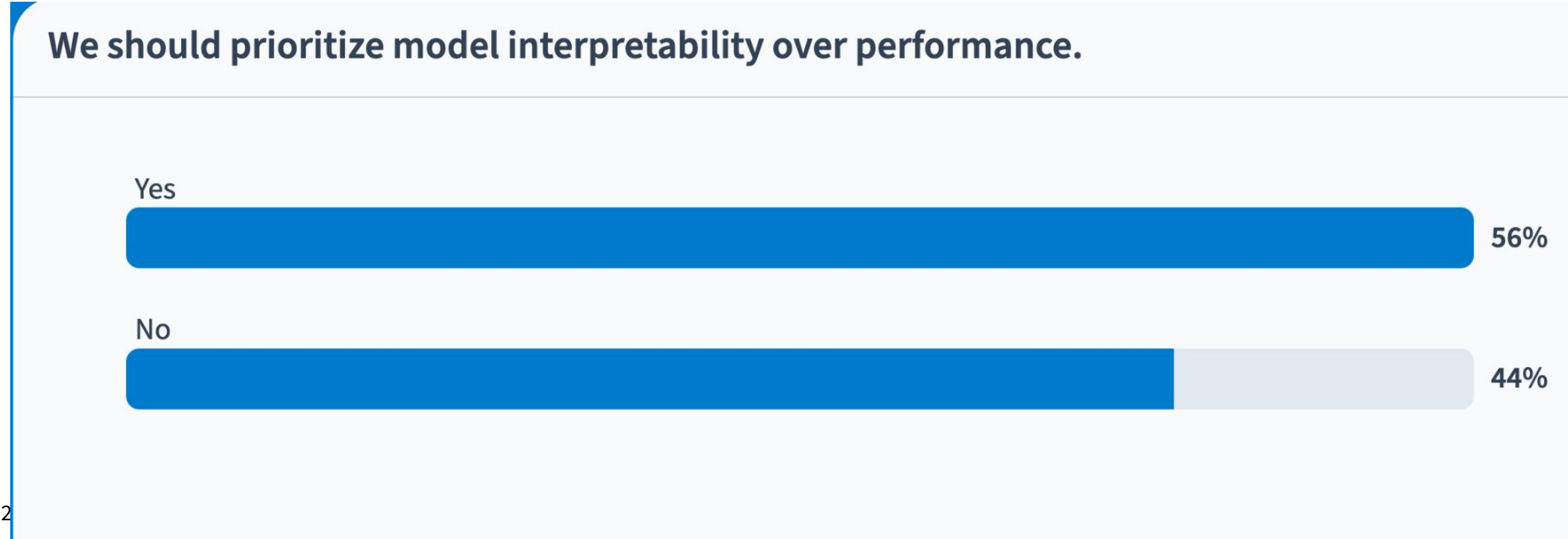
Hot-take **(BEFORE)**



We should prioritize model interpretability over performance.

Hot-take **(AFTER)**

We should prioritize model interpretability over performance.



We should prioritize model interpretability over performance.

Yes 56%

No 44%