CS 839: Foundation Models
Data

Fred Sala

University of Wisconsin-Madison

Oct. 19, 2023

Announcements

- **Logistics:**
  - Presentation Sign-up
    https://docs.google.com/spreadsheets/d/1SqXAtm6VXyofmKh0U3jaH8qg0v6nydnxoptauI8Z_1g/edit?usp=sharing
  - OH Cancelled Today ☹
- Class roadmap:

Tuesday Oct. 17	RLHF
Thursday Oct. 19	Data
Tuesday Oct. 24	Multimodal and Specialized Foundation Models
Thursday Oct. 26	Knowledge
Tuesday Oct. 31	Scaling & Scaling Laws

Outline

- **Finish RLHF**
  - Challenges, open questions, DPO variation
- **Datasets**
  - Trends, common crawl, properties, alternatives
- **Curating Datasets**
  - Filtering, Deduplication, Implications

Outline

- **Finish RLHF**
  - Challenges, open questions, DPO variation
- **Datasets**
  - Trends, common crawl, properties, alternatives
- **Curating Datasets**
  - Filtering, Deduplication, Implications

RLHF Problems

Lots of challenges!

- **Casper et al**, “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback”

- Challenges everywhere, all three phases:
  - **In human feedback,**
  - **In obtaining reward model,**
  - **In obtaining the policy**

RLHF Problems: **Human Feedback**

- Need to obtain some kind of “representative” collection of feedback providers
- **Simpler:**
  - Some people have biases
  - Mistakes due to lack of care (standard in crowdsourcing)
  - Adversarial data poisoners
- **Harder:**
  - In tough settings, what is “good” output?
  - Possible to manipulate humans

RLHF Problems: **Human Feedback**

- Additionally, **need high-quality data**.
- Expensive to hand-craft good prompts to drive feedback

- Feedback quality:
  - Tradeoffs in feedback levels
  - Ideally, rich
  - But harder to work with to train reward

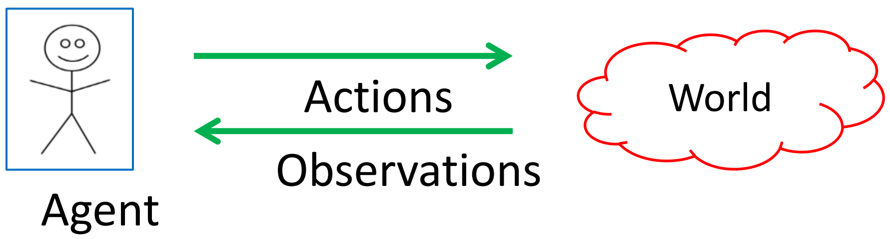
RLHF Problems: **Reward Model**

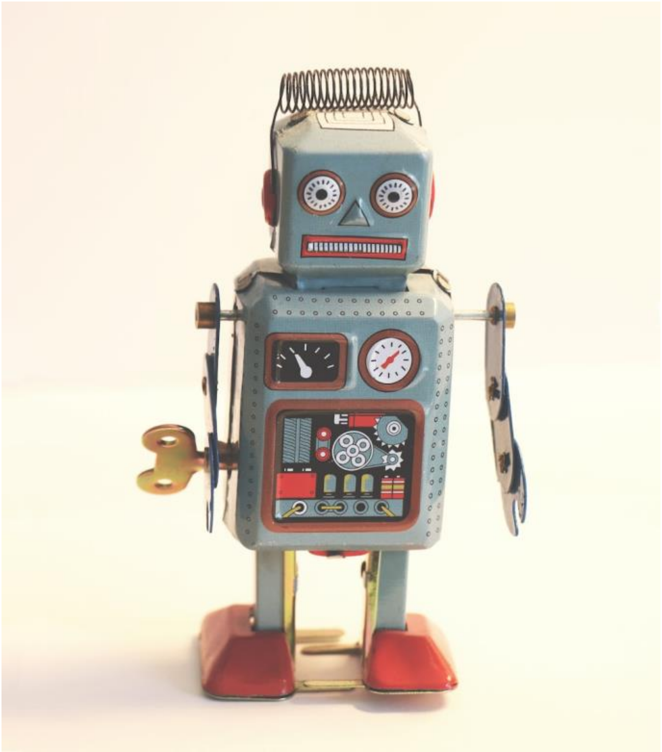
- Values can be difficult to express as a reward function
- May need to combine multiple reward functions:
  - What’s a “universal” one? People are different

- Reward Hacking
  - In tough settings, what is “good” output?
  - Possible to manipulate humans

RLHF Problems: **Training**

- The RL in RLHF can be difficult
- Also, learned policies **do not necessarily generalize to other environments**





RLHF **Alternatives**

- **Direct preference optimization** (DPO)
  - Bypass separate trained reward model: just use preference information **directly** (Rafailov et al,'23)
  - **How?** Model a preference distribution from samples, integrate into a single loss (one-stage approach)

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right].$$

  - **Gradient step:**

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \bigg] \bigg]$$

Break & Questions

Outline

- **Finish RLHF**
  - Challenges, open questions, DPO variation
- **Datasets**
  - Trends, common crawl, properties, alternatives
- **Curating Datasets**
  - Filtering, Deduplication, Implications

Trend is Generally **Bigger** and **More General**

Let's look at **GPT family training**

- **GPT1**:
  - BookCorpus: 4.5 GB 7000 unpublished books.

- **GPT2**:
  - "scraped all outbound links from Reddit … which received at least 3 karma."
  - Produced WebText, text data of 45 million links
  - "Post deduplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text"

Trend is Generally **Bigger** and **More General**

Let's look at **GPT family training**

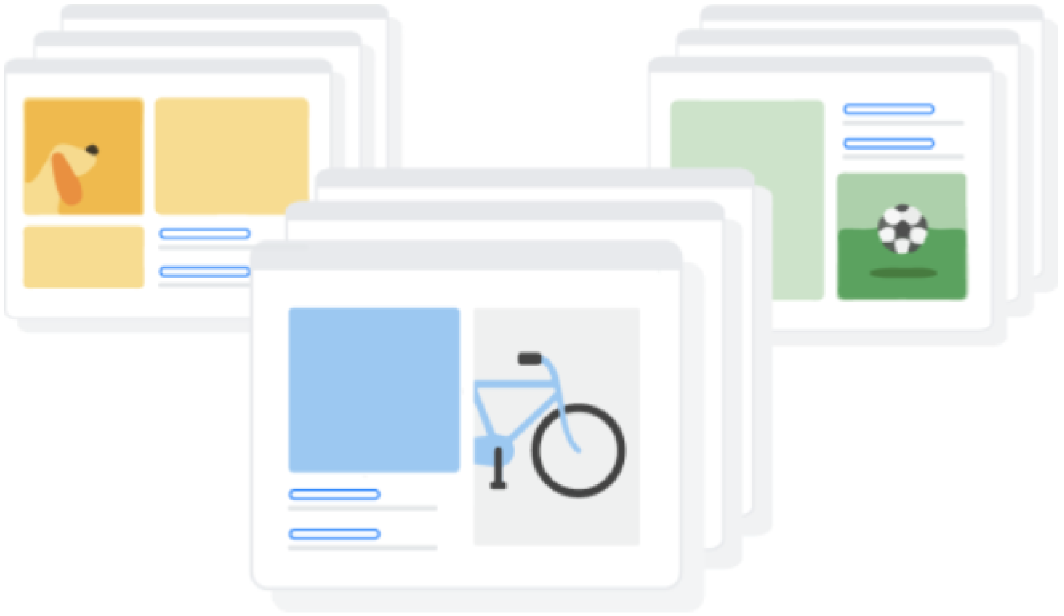
- **GPT3**:
  - A mixture of a bunch of things,

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

Brown et al '20

How Much Data Can We Get?

- One standard: Google search index
  - 100 petabytes



The Google Search index contains hundreds of billions of webpages and is well over 100,000,000 gigabytes in size. It's like the index in the back of a book — with an entry for every word seen on every webpage we index. When we index a webpage, we add it to the entries for all of the words it contains.

https://www.google.com/search/howsearchworks/how-search-works/organizing-information/

Common Crawl

- Organization that crawls web and releases snapshots
  - Still orders of magnitude below Google
  - But really big!

Crawl date	Size in TiB	Billions of pages	Comments
June 2023	390	3.1	Crawl conducted from May 27 to June 11, 2023
April 2023	400	3.1	Crawl conducted from March 20 to April 2, 2023
February 2023	400	3.15	Crawl conducted from January 26 to February 9, 2023
December 2022	420	3.35	Crawl conducted from November 26 to December 10, 2022
October 2022	380	3.15	Crawl conducted in September and October 2022

https://commoncrawl.org/

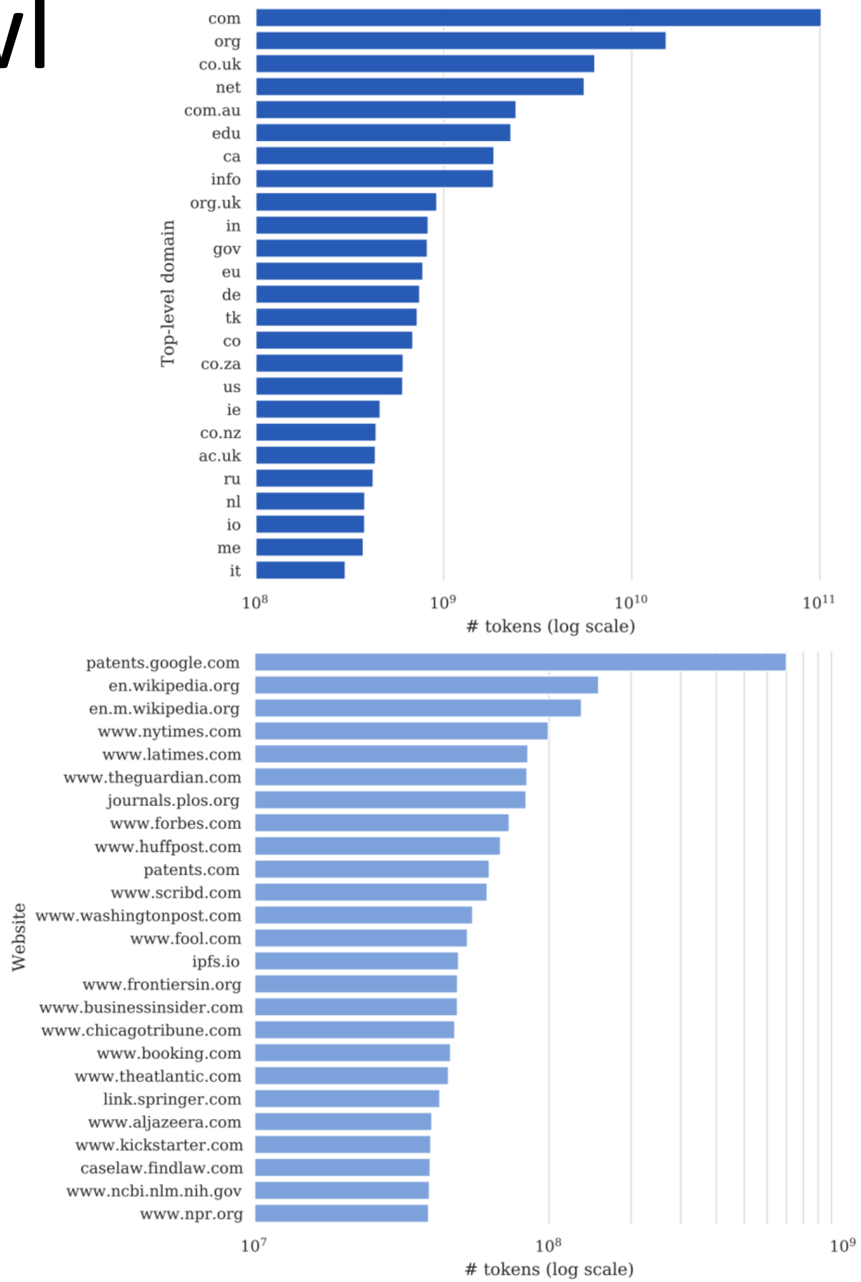
Some Issues…

- Lots of data, but
  - Not representative!
  - Basically who is on the Internet most: younger users, developed nations
  - Tracking **composition** is a key idea
  - Avoiding toxic text as well:
    - OpenWebText 2-4% of text is largely toxic (Gehman et al '20)
    - More in a later lecture

Cleaning Up Common Crawl

- **Colossal Clean Crawled Corpus** (C4)
  - Removes bad words
  - Removes code
  - Language detection
  - ~800 GB (150 billion tokens)

  - Used to train T5 (Raffel et al ’23)
  - Analyzed by Dodge et al ‘21



Dodge et al ‘21

More Issues: Contamination

- Lots of data, but
  - Leakage/contamination
  - Want our benchmarks to not have shown up in our training data
  - This is really hard to control!
    - Both inputs and outputs to benchmark tasks are there (2% to 25%)
    - Even just input can hurt

Other Places to Get Data

- **The Pile**
  - Large dataset composed of many smaller but **high-quality** parts
  - Gao et al '20 / Eleuther AI
  - Comparisons show that a lot of this data isn't covered well in crawls

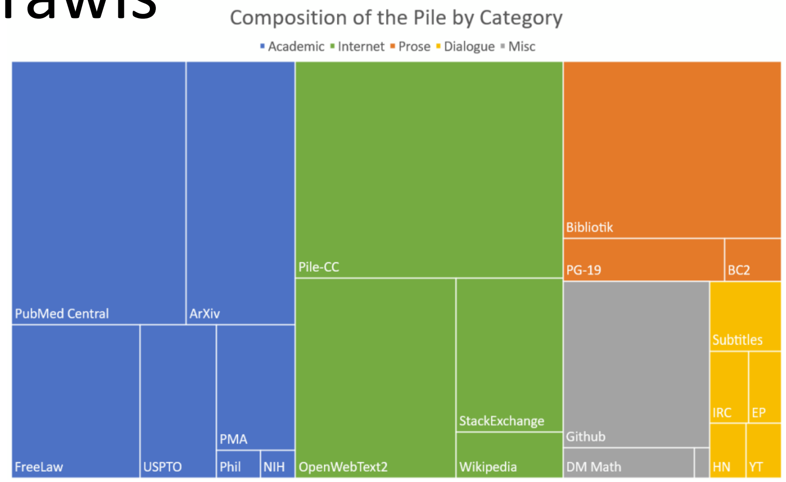


Figure 1: Treemap of Pile components by effective size.

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

Break & Questions

Outline

- **Finish RLHF**
  - Challenges, open questions, DPO variation
- **Datasets**
  - Trends, common crawl, properties, alternatives
- **Curating Datasets**
  - Filtering, Deduplication, Implications

Processing Data: **Filtering**

- As we saw, have to process data first
  - Filter out some points (toxicity, mismatch, etc)
  - Generally, we want "better" datasets
    - More diversity,
    - Less repeats.
- New benchmarks target this setting,
  - Fix the training procedure
  - Vary the data

DataComp

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

https://www.datacomp.ai/

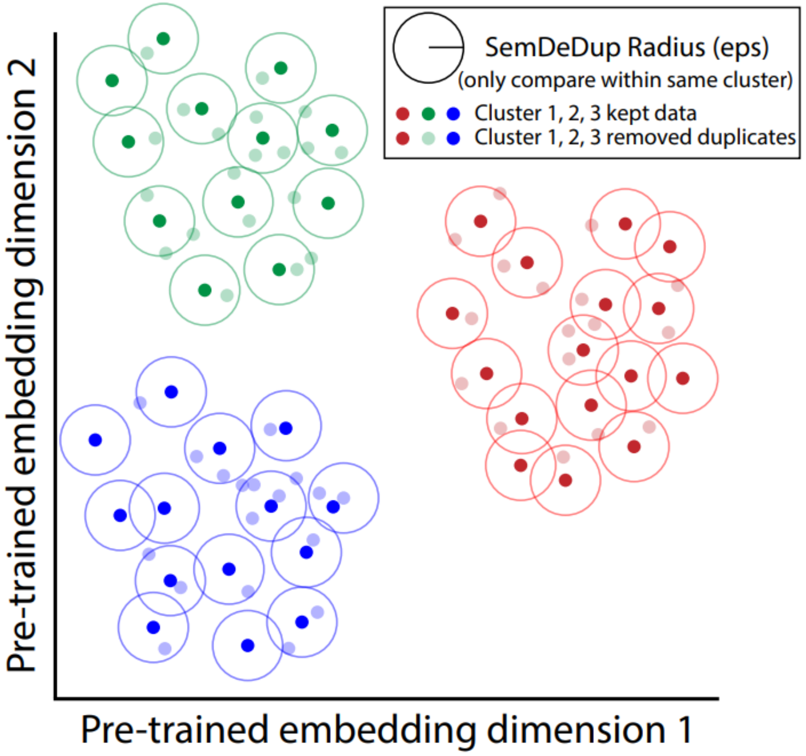
Processing Data: **Deduplication**

- "Deduplicating Training Data Makes Language Models Better ": Lee et al '22
  - Various ways to deduplicate data
  - Exact string matching
  - Approximate (hash-based, equivalent to embedding-based)

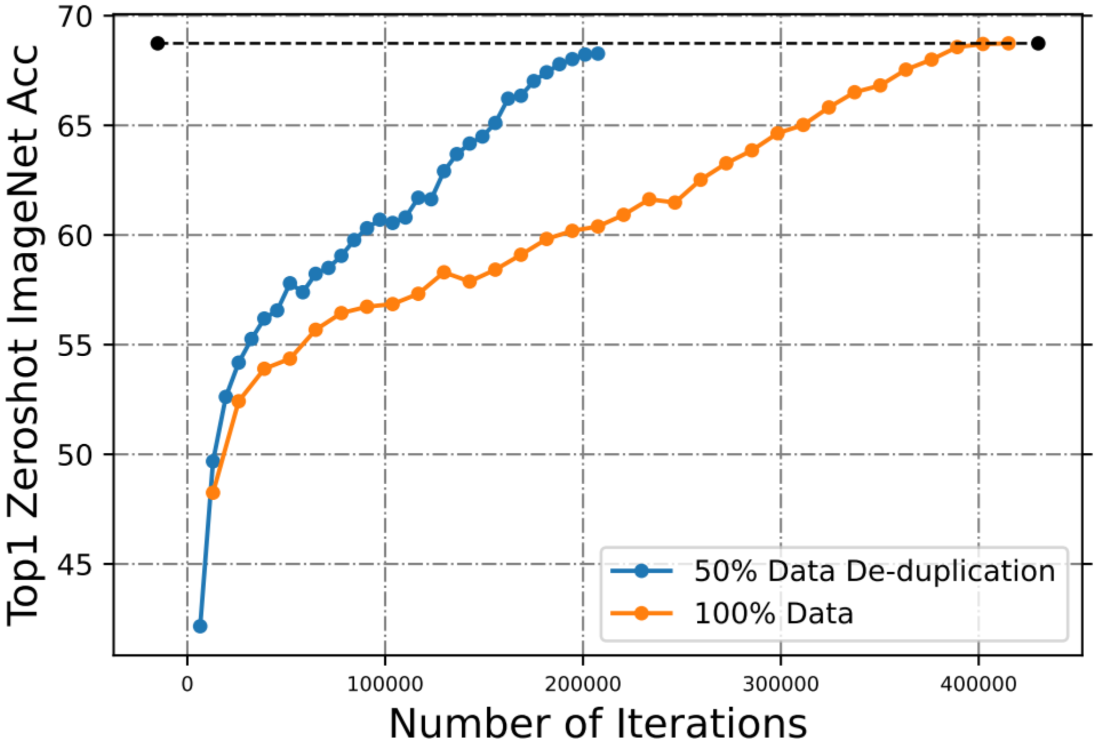
- One sentence shows up in **C4 60,000 times!**
  - `"by combining fantastic ideas, interesting arrangements, and follow the current trends in the field of that make you more inspired and give artistic touches. We'd be honored if you can apply some or all of these design in your wedding. believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!"`

Processing Data: **Semantic Deduplication**

- How to define “duplicated” for data?
  - Idea: SemDeDup uses embeddings to identify near duplicates



Abbas et al ‘23



Bibliography

- Casper et al: Stephen Casper, Xander Davies, and many others, “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback” (https://arxiv.org/abs/2307.15217)
- Rafailov et al: Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, “Direct Preference Optimization: Your Language Model is Secretly a Reward Model” (https://arxiv.org/abs/2305.18290)
- Commoncrawl.org
- Gehman et al: Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, Noah A. Smith, “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models” (https://arxiv.org/abs/2009.11462)
- Raffel et al: Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” (https://arxiv.org/abs/1910.10683)
- Dodge et al: Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, Matt Gardner, “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus” (https://arxiv.org/abs/2104.08758)
- Gao et al: Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy, “The Pile: An 800GB Dataset of Diverse Text for Language Modeling” (https://arxiv.org/abs/2101.00027)
- https://www.datacomp.ai/index.html#home
- Lee et al: Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, Nicholas Carlini, “Deduplicating Training Data Makes Language Models Better” (https://aclanthology.org/2022.acl-long.577/)
- Abbas et al: Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, Ari S. Morcos, “SemDeDup: Data-efficient learning at web-scale through semantic deduplication” (https://arxiv.org/abs/2303.09540)

Thank You!