# CS 839: Foundation Models
## **Data**

Fred Sala

University of Wisconsin-Madison

**Oct. 22, 2024**

# Announcements

- **Logistics:**
  - HW 2: Deadline **Pushed Back to Oct. 31**
  - Continue signing up for presentations
  - Project information is out
- Class roadmap:

| Thursday Oct. 19 | Data |
| --- | --- |
| Tuesday Oct. 24 | Evaluation |
| Thursday Oct. 26 | Multimodal models |
| Tuesday Oct. 31 | Diffusion Models |
| Tuesday Nov. 5 | Scaling & Scaling Laws |

# Outline

- **Pretraining Datasets**
  - Trends, common crawl, properties, alternatives
- **Other Datasets**
  - Instruction-tuning data, Reward model-type data
- **Curating Data**
  - Filtering, Deduplication, Implications

# Outline

- **Pretraining Datasets**
  - Trends, common crawl, properties, alternatives
- **Other Datasets**
  - Instruction-tuning data, Reward model-type data
- **Curating Data**
  - 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  - BookCorpus: 4.5 GB 7000 unpublished books.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**

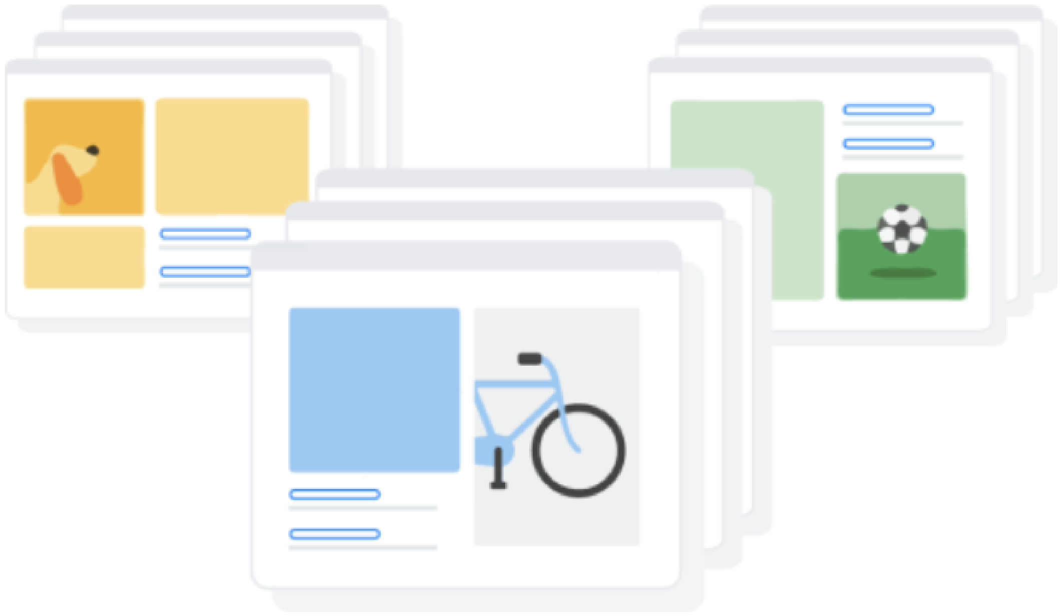
- **GPT3**:
  - A mixture of a bunch of things,

| 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…

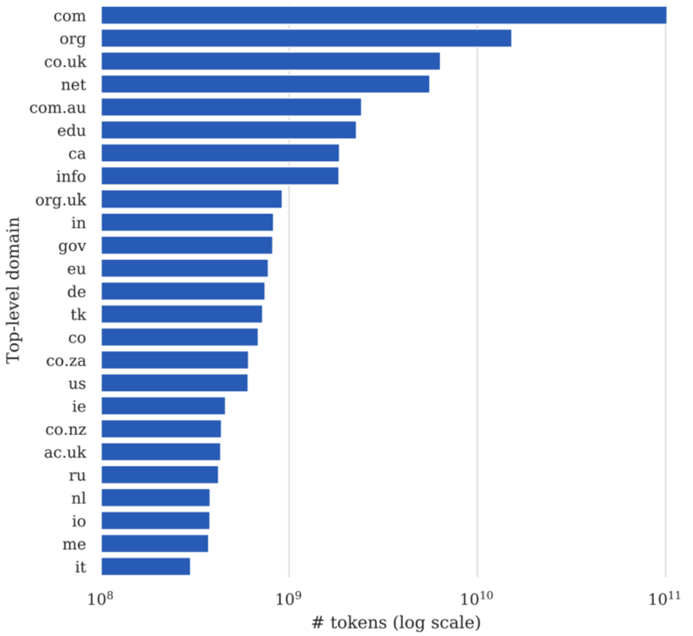
- Lots of data, but
  - Not representative!
  - Basically who is on the Internet most: younger users, developed nations
  - Tracking **composition** is a key idea

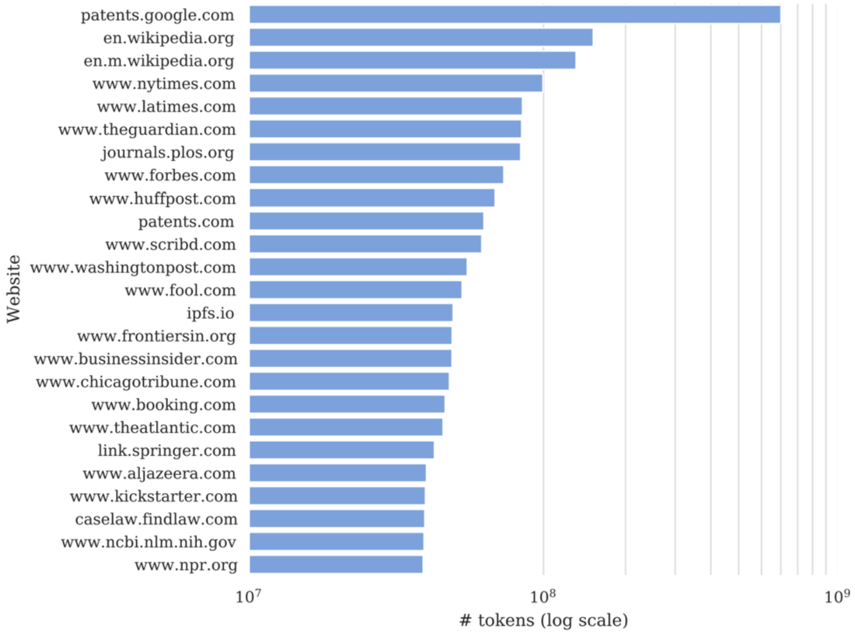
  - Avoiding toxic text as well:
    - OpenWebText 2-4% of text is largely toxic (Gehman et al '20)
    - More in a later lecture

# 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

- **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

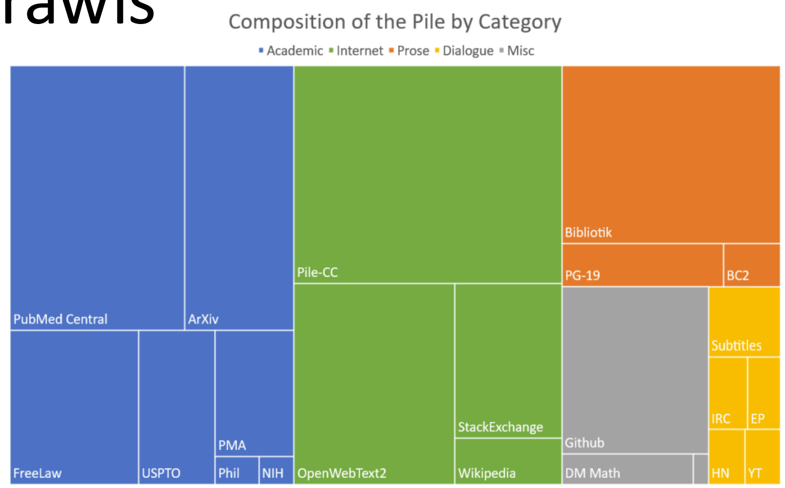



Figure 1: Treemap of Pile components by effective size.

| 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** |

# Other Places to Get Data: RedPajama

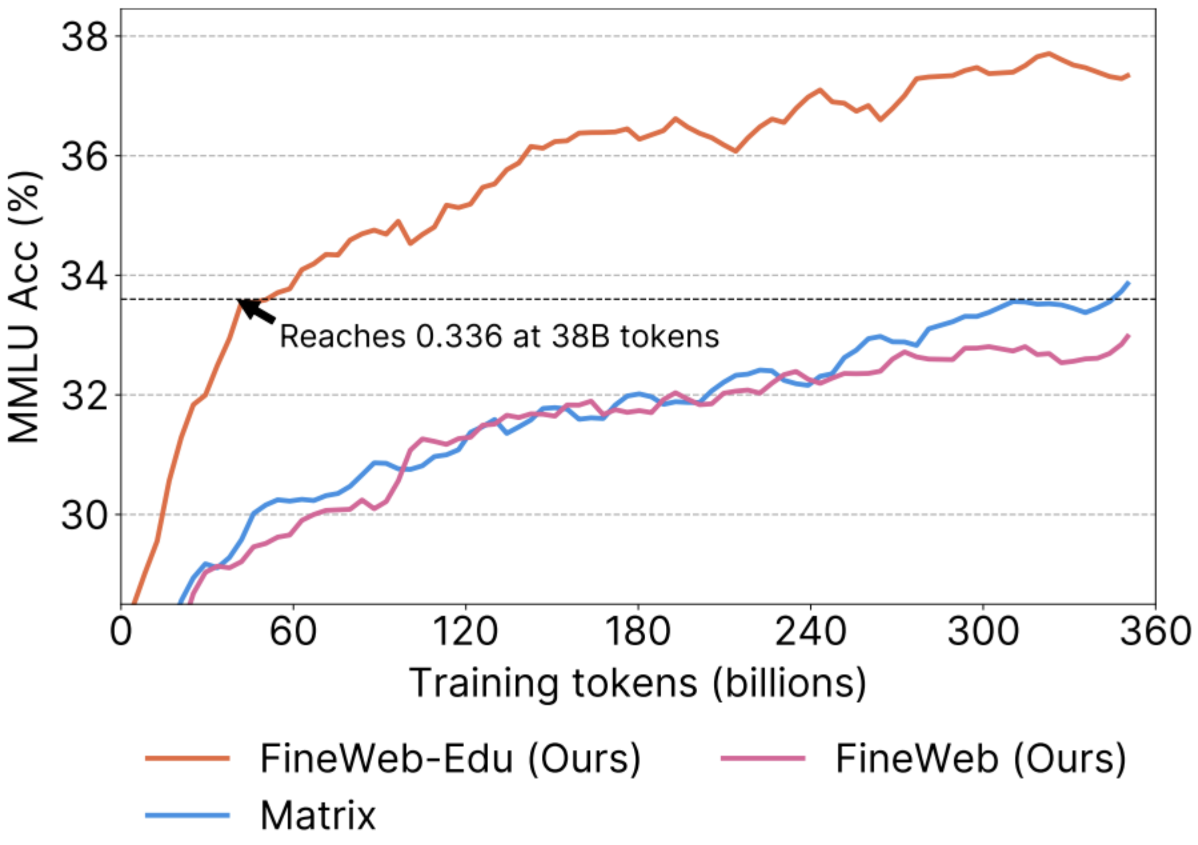
- **RedPajama v2**
  - Open dataset with 30 trillion tokens
  - Oct ’23 / Together AI
  - Pre-computed quality annotations
    - “ML classifiers on data quality, minhash results that can be used for fuzzy deduplication, or heuristics such as “the fraction of words that contain no alphabetical character”. ”

| | # Documents | Estimated Token count (deduped) |
|---|---|---|
| en | 14.5B | 20.5T |
| de | 1.9B | 3.0T |
| fr | 1.6B | 2.7T |
| es | 1.8B | 2.8T |
| it | 0.9B | 1.5T |
| Total | 20.8B | 30.4T |

github.com/togethercomputer/RedPajama-Data

# Other Places to Get Data: FineWeb

- **FineWeb**
  - Open dataset with 15 trillion tokens
  - June ’24 / Hugging Face
  - Additional filtered ”educational” data
  - Full data construction and experimental details available.



# Other Places to Get Data: Synthetic?

- **Can create synthetic data for all phases of training…**
  - Typically easier for particular domains
  - And for instruction tuning / fine-tuning / alignment

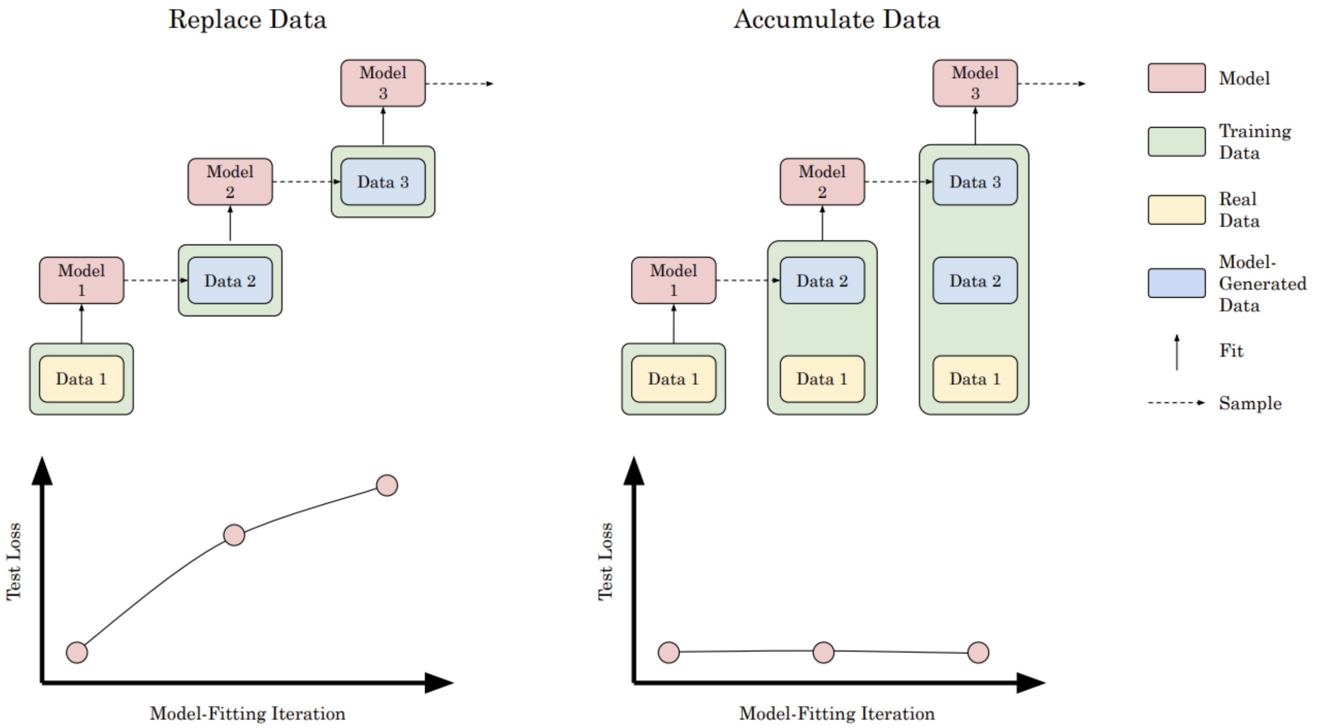
## Best Practices and Lessons Learned on Synthetic Data for Language Models

**Ruibo Liu[1], Jerry Wei[1], Fangyu Liu[1], Chenglei Si[2], Yanzhe Zhang[3], Jinmeng Rao[1], Steven Zheng[1], Daiyi Peng[1], Diyi Yang[2], Denny Zhou[1] and Andrew M. Dai[1]**
[1]Google DeepMind, [2]Stanford University, [3]Georgia Institute of Technology

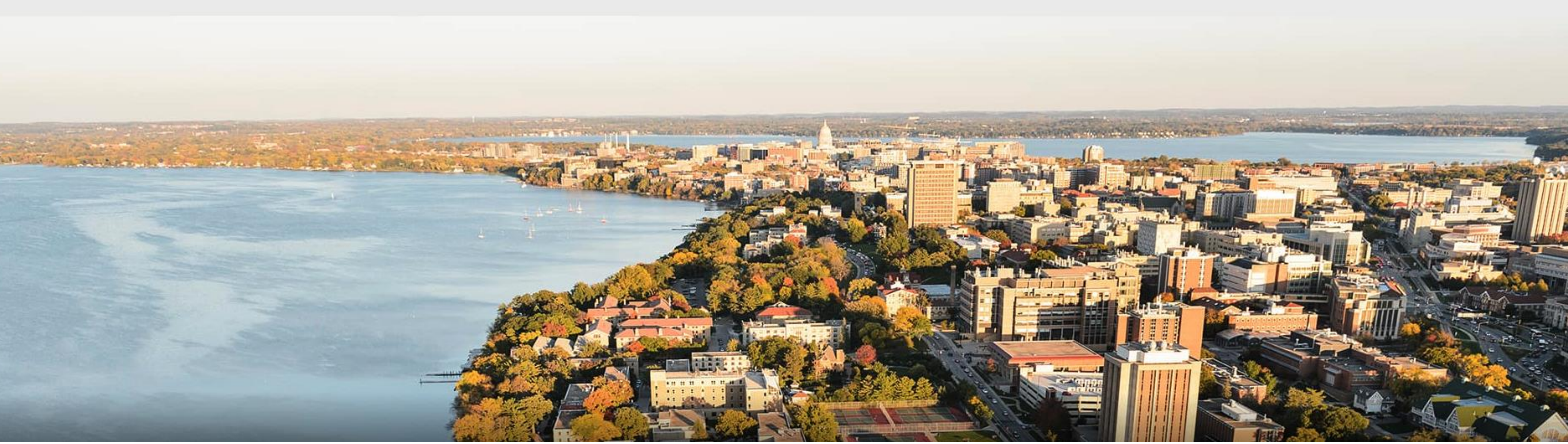
**The success of AI models relies on the availability of large, diverse, and high-quality datasets, which can be challenging to obtain due to data scarcity, privacy concerns, and high costs. Synthetic data has emerged as a promising solution by generating artificial data that mimics real-world patterns. This paper provides an overview of synthetic data research, discussing its applications, challenges, and future directions. We present empirical evidence from prior art to demonstrate its effectiveness and highlight the importance of ensuring its factuality, fidelity, and unbiasedness. We emphasize the need for responsible use of synthetic data to build more powerful, inclusive, and trustworthy language models.**

# Other Places to Get Data: Synthetic?

- **One risk: possibility of model collapse**
  - Idea: feeding data back into model for training just causes the model to collapse into a single
  - Solution: reinforce/accumulate some real data




Gerstgrasser et al ‘24



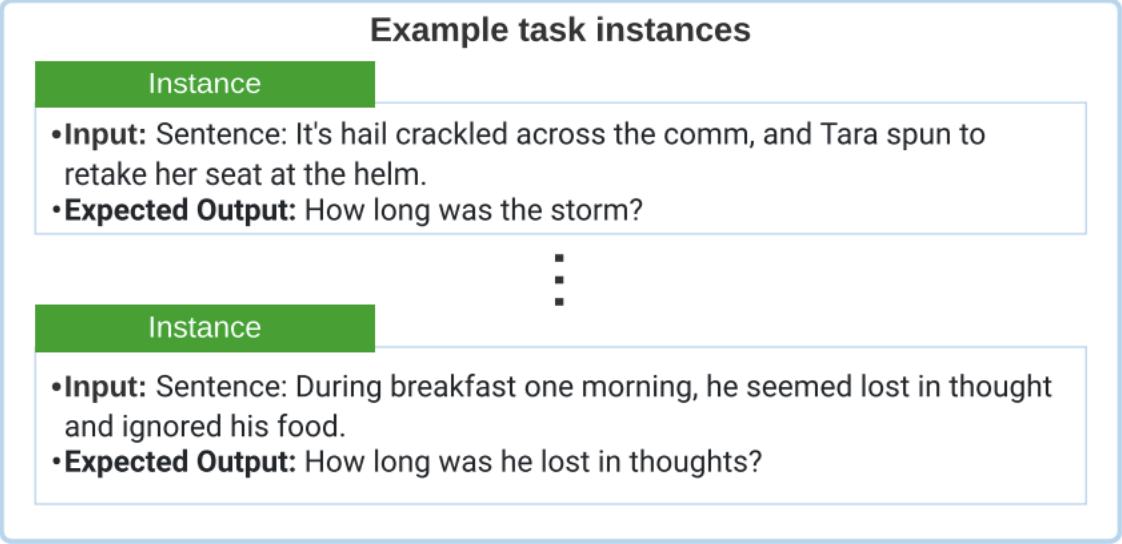
# Break & Questions

# Outline

- **Pretraining Datasets**
  - Trends, common crawl, properties, alternatives
- **Other Datasets**
  - Instruction-tuning data, Reward model-type data
- **Curating Data**
  - Filtering, Deduplication, Implications

# Other Forms of Data: Instruction Tuning

- **Natural Instructions**
  - Open dataset
  - Mishra et al, ‘22
  - 61 tasks, ~200K instructions
    - Note: scale much smaller than pretraining



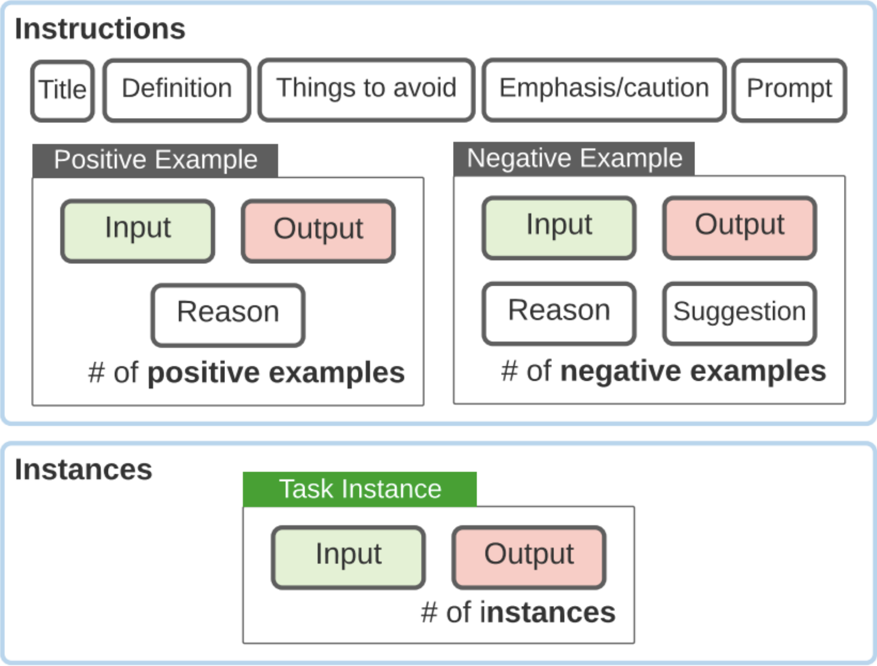
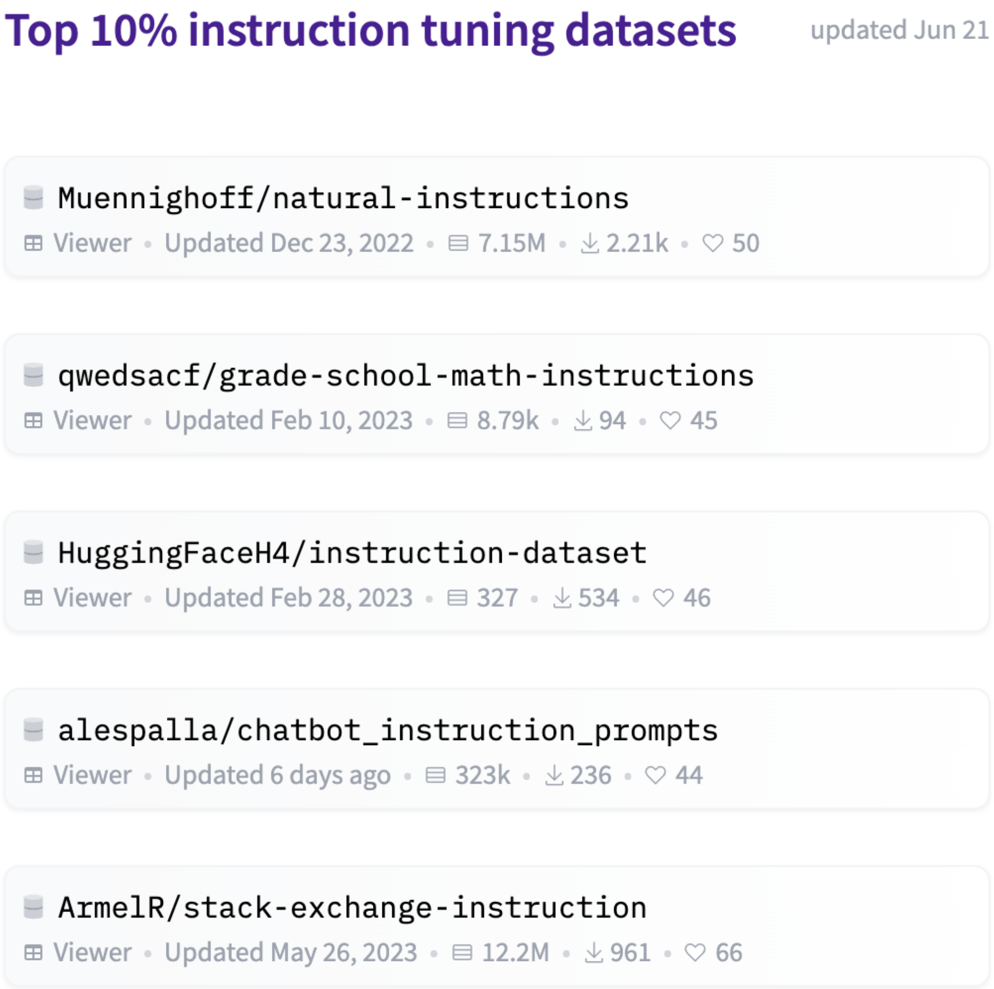




Figure 4: The schema used for representing instruction in NATURAL INSTRUCTIONS (§4.1), shown in plate notation.

# Other Forms of Data: Instruction Tuning

- **Lots more available,**
  - Hugging face has a great collection

- Pick out ones suitable for your desired instruction-tuned model




huggingface.co

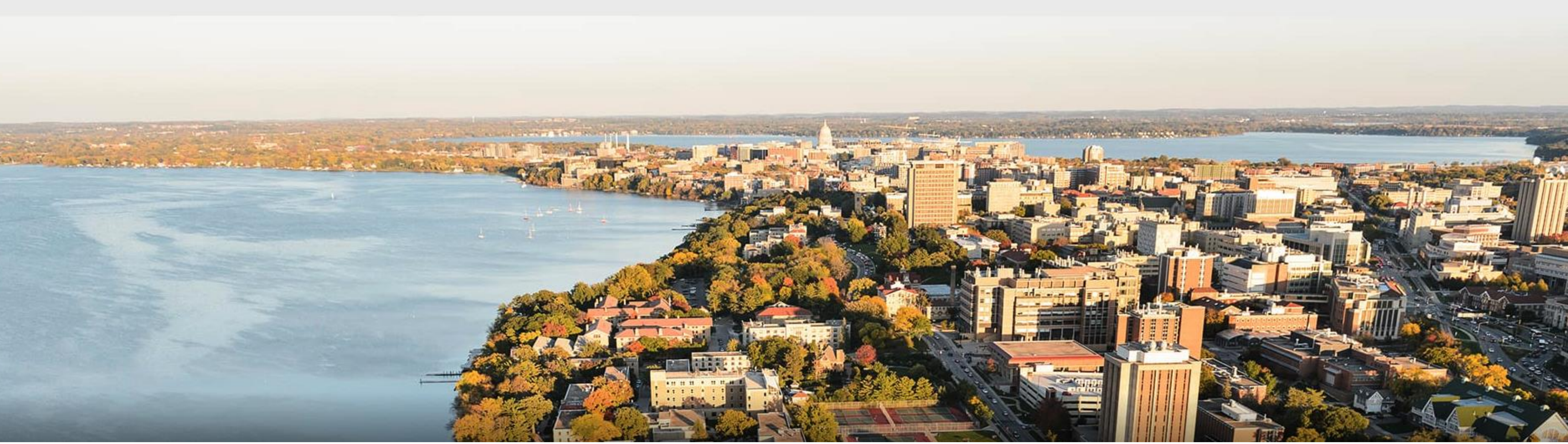
# Other Forms of Data: Alignment Data

- **HelpSteer, HH RLHF, etc.**
  - Often annotated with attributes to help alignment

| *Name* | *Helpfulness-relevant Attributes* | *N conv. (k)* | Mean Length in chars (Std.) Prompt | Response |
|---|---|---|---|---|
| HELPSTEER | Helpfulness, Correctness, Coherence, Complexity, Verbosity | 37.1 | 2491.8 (1701.7) | 497.3 (426.7) |
| Open Assistant | Quality, Creativity, Humor | 59.4 | 397.5 (620.8) | 396.2 (618.8) |
| HH RLHF | - | 337.7 | 794.4 (706.9) | 310.7 (311.4) |

Table 1: Overview of Open-source Helpfulness Preference Modeling Datasets

Wang et al ‘23



# Break & Questions

# Outline

- **Pretraining Datasets**
  - Trends, common crawl, properties, alternatives
- **Other Datasets**
  - Instruction-tuning data, Reward model-type data
- **Curating Data**
  - 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-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

nark where the models are fixed and the

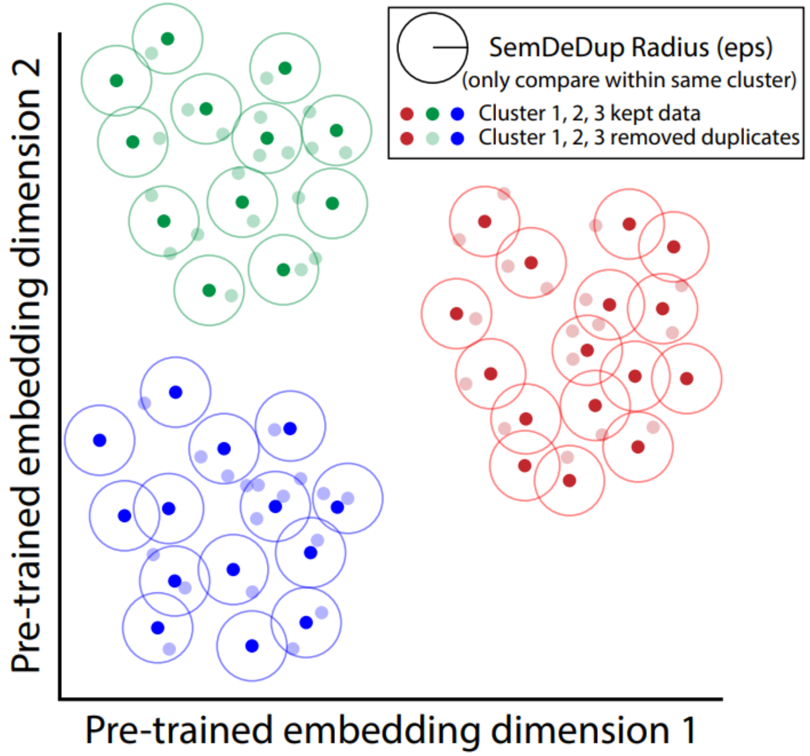
# Processing Data: **Deduplication**

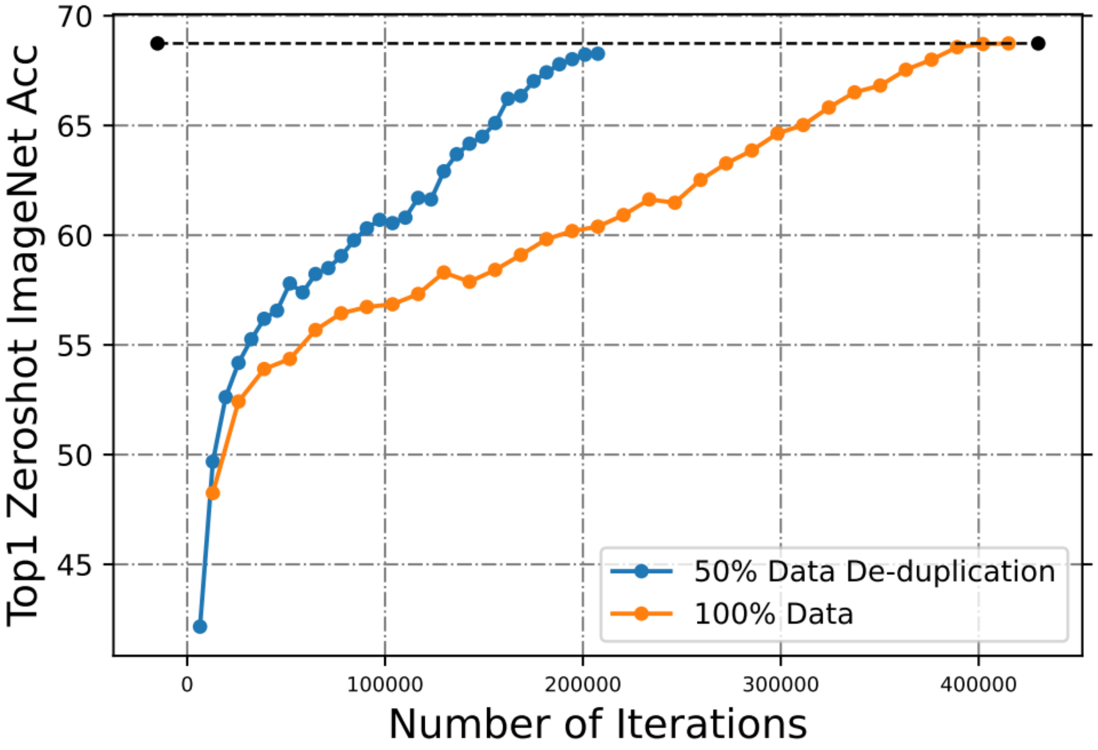
- “Deduplicating Training Data Makes Language Models Better “: Lee et al ’22
  - Various ways to deduplicate data
  - Exact string matching
  - Approximate (hash-based, equivalent to embedding-based)

- 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


- 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://github.com/togethercomputer/RedPajama-Data
- Penedo et al: Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf , "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale" (https://arxiv.org/abs/2406.17557)
- Liu et al: Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, Andrew M. Dai, "Best Practices and Lessons Learned on Synthetic Data" (https://arxiv.org/abs/2404.07503)
- Gerstgrasser et al: Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, Sanmi Koyejo, "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data" (https://arxiv.org/abs/2404.01413)
- Mishra et al '22: Swaroop Mishra, Daniel Khashabi, Chitta Baral, Hannaneh Hajishirzi, "Cross-Task Generalization via Natural Language Crowdsourcing Instructions" (https://arxiv.org/abs/2104.08773)
- Wang et al '23: Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, Oleksii Kuchaiev, " HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM" (https://arxiv.org/abs/2311.09528)
- 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!