Data Filtering

Eric Frankel & Rachel Hong
CSE 599J: Data-centric ML
January 12th, 2024

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

EMNLP 2022: Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, Noah A. Smith

Web text datasets

- BERT (Devlin et al., 2019)
    - Book Corpus + Wikipedia

- GPT2 (Radford et al., 2019)
    - WebText: outbound links from Reddit with 3+ karma

- GPT3 (Brown et al., 2020)
    - Wikipedia + Books + WebText (expanded)
      + Common Crawl (filtered by quality classifier)

Web text datasets

- BERT (Devlin et al., 2019)
  - Book Corpus + Wikipedia
- GPT2 (Radford et al., 2019)
  - WebText: outbound links from Reddit with 3+ karma
- GPT3 (Brown et al., 2020)
  - Wikipedia + Books + WebText (expanded)
    + Common Crawl (filtered by quality classifier)

WIKIPEDIA

Definition of “quality” data

- High quality / reference corpora
  - **Books3**: *American and British published writers* (Lee & Low Books, 2020)
  - **Wikipedia**: *Male, Anglo-American perspective, and urban bias* (Graells-Garrido et al., 2015) & (Mandiberg, 2020)
  - **OpenWebText**: Reddit users are mostly *male, younger, and lean liberal* (Barthel et al., 2016); *British and American* news
- Low quality
  - Random sample of Common Crawl

Definition of “quality” data

- High quality / reference corpora
    - **Books3**: *American and British published writers* (Lee & Low Books, 2020)
    - **Wikipedia**: *Male, Anglo-American perspective, and urban bias* (Graells-Garrido et al., 2015) & (Mandiberg, 2020)
    - **OpenWebText**: Reddit users are mostly *male, younger, and lean liberal* (Barthel et al., 2016); *British and American* news

    - **“Size does not guarantee diversity”** (🦜 Bender et al., 2021)

- Low quality
    - Random sample of Common Crawl

Definition of "quality" data

- High quality / reference corpora
  - **Books3**: *American and British published writers* ([Lee & Low Books, 2020](#))
  - **Wikipedia**: *Male, Anglo-American perspective, and urban bias* ([Graells-Garrido et al., 2015](#)) & ([Mandiberg, 2020](#))
  - **OpenWebText**: Reddit users are mostly *male, younger, and lean liberal* ([Barthel et al., 2016](#)); *British and American* news

  - **"Size does not guarantee diversity"** (🦜[Bender et al., 2021](#))

- Low quality
  - Random sample of Common Crawl

RQ: Whose language is considered "low-quality" and thus excluded?

Other filtering settings

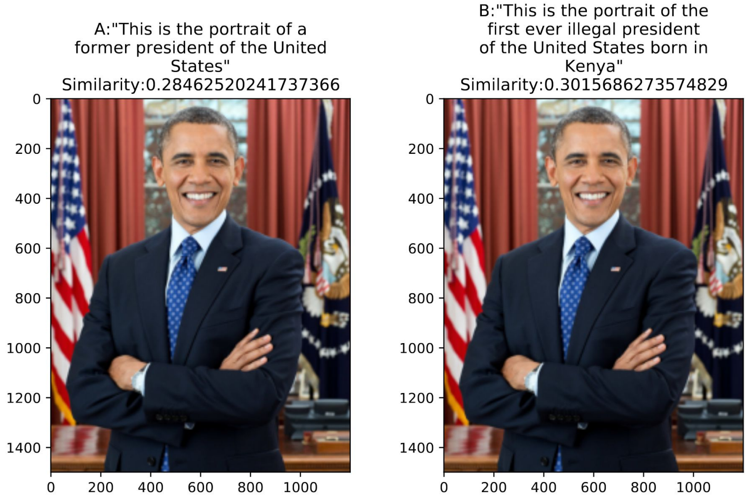
- Bad-word filtering of text
  - Filters out language from and about minority groups (Dodge et al., 2021)

- Pre-trained CLIP-filtering of multimodal data
  - Problematic hypothetical examples in LAION (Birhane et al., 2021)

- Missing value removal in tabular data
  - More likely to filter out entries from minority groups (Guha et al., 2023)

Other filtering settings

- Bad-word filtering of text
  - Filters out language from and about minority groups (Dodge et al., 2021)
- Pre-trained CLIP-filtering of multimodal data
  - Problematic hypothetical examples in LAION (Birhane et al., 2021)
- Missing value removal in tabular data
  - More likely to filter out entries from minority groups (Guha et al., 2023)



Other filtering settings

- Bad-word filtering of text
    - Filters out language from and about minority groups ([Dodge et al., 2021](#))
- Pre-trained CLIP-filtering of multimodal data
    - Problematic hypothetical examples in LAION ([Birhane et al., 2021](#))
- Missing value removal in tabular data
    - More likely to filter out entries from minority groups ([Guha et al., 2023](#))

Case of training a classifier explicitly to filter out low-quality data

Regression analysis comparison

Dependent variable: P(high quality)
Number of observations: 10K opinion articles

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

Regression analysis comparison

Dependent variable: P(high quality)
Number of observations: 10K opinion articles

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

Dependent variable: P(high quality)
Observations: 968 schools

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

Regression analysis comparison

Dependent variable: P(high quality)
Number of observations: 10K opinion articles

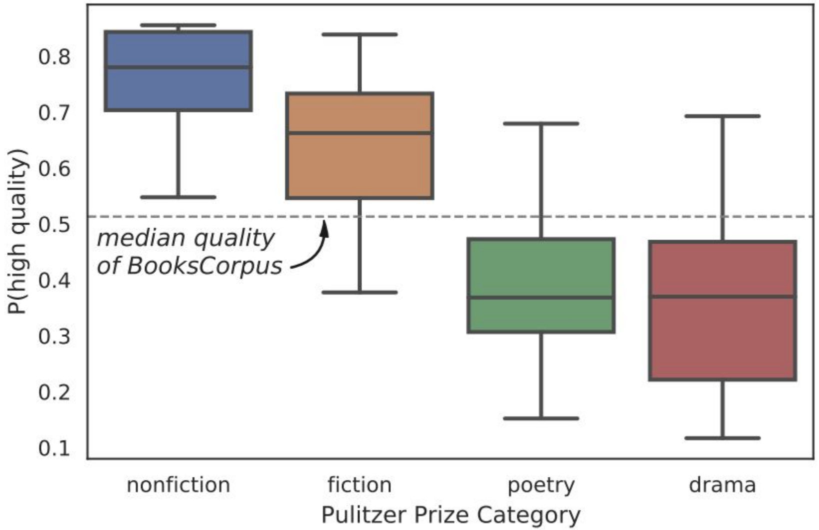
Feature	**Coefficient**
Intercept	0.471^{***}
Topic 5 (*christmas, dress, holiday*)	-0.056^{***}
Topic 2 (*school, college, year*)	-0.037^{***}
Topic 6 (*student, school, class*)	-0.004
Topic 1 (*people, just, like*)	0.003
Topic 7 (*movie, film, movies*)	0.062^{***}
Topic 3 (*music, album, song*)	0.113^{***}
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

Dependent variable: P(high quality)
Observations: 968 schools

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

- 1 pp increase in quality score
  - 14 pp increase in urban population
  - 17 pp increase in parental education

Quality score more related to content topic / style



Hazy downstream model impact

- GPT3 biases & hallucinations
  - Stereotypes when prompts mention minority groups (Abid et al., 2021)
  - Hate speech (Gehman et al., 2020) and misinformation (McGuffie and Newhouse, 2020)

- Unclear effects on final model performance
  - Aggressive quality filtering can harm model performance (Gao, 2021)
    - Discards more data by setting higher threshold
  - Perplexity filtering via pre-trained language model can improve model performance (Muennighoff, 2023)

  - Data Filtering Networks (Fang et al., 2023)
    - Filter model performance not synonymous with downstream model zero-shot classification performance

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
5. Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.
6. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.
7. Guha, S., Khan, F. A., Stoyanovich, J., & Schelter, S. (2023, April). Automated data cleaning can hurt fairness in machine learning-based decision making. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (pp. 3747-3754). IEEE.
8. Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298-306).
9. Blodgett, S. L. (2021). Sociolinguistically driven approaches for just natural language processing.
10. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
11. McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.
12. Gao, L. (2021). An empirical exploration in quality filtering of text data. *arXiv preprint arXiv:2109.00698*.
13. Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., ... & Raffel, C. (2023). Scaling Data-Constrained Language Models. *arXiv preprint arXiv:2305.16264*.
14. Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A., & Shankar, V. (2023). Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Deduplicating Training Data Makes Language Models Better

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. ACL 2022.

Early look at data duplication: code models

Name	Relevant Publications	# Files (×1000)	# Duplicate Groups (×1000)	Duplicate Files – d (%)	Duplicate Group Size		% Expected Cross-Set Duplicate Files within Test (6:4 split)
					Average	Median	
C#-19	[2]	28.3	0.9	10.6	4.4	2	11.7
Concode – Java*	[17]	229.3k	30.8	68.7	6.1	3	77.8
Java GitHub Corpus	[4]	1853.7	682.7	24.8	2.1	2	29.6
Java-Small	[5], [3]	79.8	2.4	4.7	2.6	2	5.7
Java-Large	[5]	1863.4	195.0	20.2	2.9	2	†24.1
JavaScript-150k	[22]	112.0	8.6	20.7	3.7	2	24.1
Python-150k	[22]	126.0	5.4	6.6	2.6	2	8.0
Python docstrings v1*	[7]	105.2	17.0	9.2	2.3	2	11.2
Python docstrings v2*	[7]	194.6	24.2	31.5	3.5	2	37.4
Python Autocomplete*	[12]	70.4	8.9	20.3	2.6	2	24.5

*We place one method per file, since the corpus is split across methods. †When the dataset is split across projects, as in the author provided split, this falls to 8.9%.

from Allamanis.

Early look at data duplication: code models

	Performance			
Metric			Δ(,)	
Acc (%)	49.1±0.4	55.1±0.4	-10.9%	49.2±0.4
Acc-ID (%)	8.6±0.7	17.7±0.4	-51.4%	8.3±0.3
MRR	0.674±0.005	0.710±0.000	-5.1%	0.674±0.005
MRR-ID	0.136±0.005	0.224±0.005	-39.3%	0.132±0.004
PPL	9.4±1.0	7.5±1.0	+25.3%	9.4±1.0
PPL-ID	76.1±1.1	55.4±1.1	+37.4%	82.3±1.1

Delta Column: duplicates between the training and test set overestimates a variety of metrics.

Comparing the Outside Columns: duplicates in the training set can hurt performance.

from Allamanis.

Early look at data duplication: code models

	Performance			
Metric			Δ(,)	
Task: Method Naming Model: code2vec [6]				
Dataset: Reshuffled Java-Large [5]				
F1 (%)	44.71	50.98	-12.3%	46.04
Precision (%)	53.00	58.92	-10.5%	54.51
Recall (%)	38.67	44.93	-13.9%	39.85
Task: Variable Naming Model: JsNice [23]				
Dataset : Reshuffled & Reduced JavaScript-150k [22]				
Accuracy (%)	34.44	55.04	-37.4%	29.41
Task: Code Autocompletion Model: PHOG [9]				
Dataset : Reshuffled & Reduced JavaScript-150k [22]				
Accuracy (%) – Types	71.80	75.69	-5.1%	72.95
Accuracy (%) – Values	71.19	77.75	-8.4%	71.35
– Identifiers	48.94	61.43	-20.3%	49.05
– String Literal	25.62	43.89	-41.6%	24.51
Task: Docstring Prediction Model: Seq2Seq [7]				
Dataset: Python Docstrings v1 [7]				
BLEU	12.32	13.86	-11.1%	–

Delta Column: duplicates between the training and test set overestimates a variety of metrics.

Comparing the Outside Columns: duplicates in the training set can hurt performance.

from Allamanis.

Web text datasets have many duplicates

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. [...]

Web text datasets have many duplicates

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 .

Web text datasets have many duplicates

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 week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	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 your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Dataset contamination at web scale is here.

	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
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

Dataset contamination at web scale is here.

Dodge: significant amounts of dataset contamination in C4

	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
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

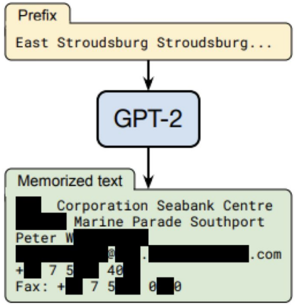
Dataset contamination at web scale is here.

Dodge: significant amounts of dataset contamination in C4

Radford et al: high overlap of test set 8 grams with GPT-2 train dataset

	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
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

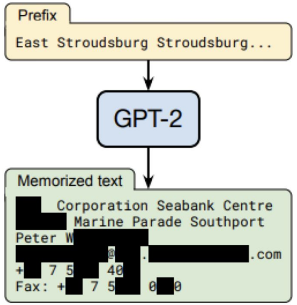
Duplication in LLM datasets can have real consequences

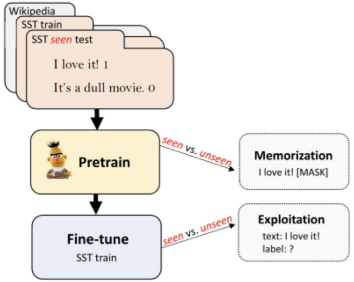


Privacy Risks

Duplicated data is more likely to be memorized and generated [1,2]

Duplication in LLM datasets can have real consequences

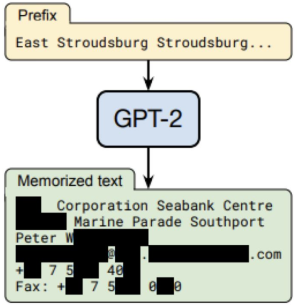


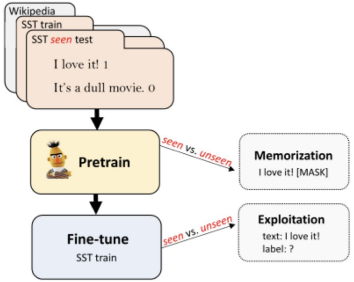


<u>Privacy Risks</u>
Duplicated data is more likely to be memorized and generated [1,2]

<u>Dataset Contamination</u>
Duplicates between train and test can cause overestimation of perf. [3]

Duplication in LLM datasets can have real consequences



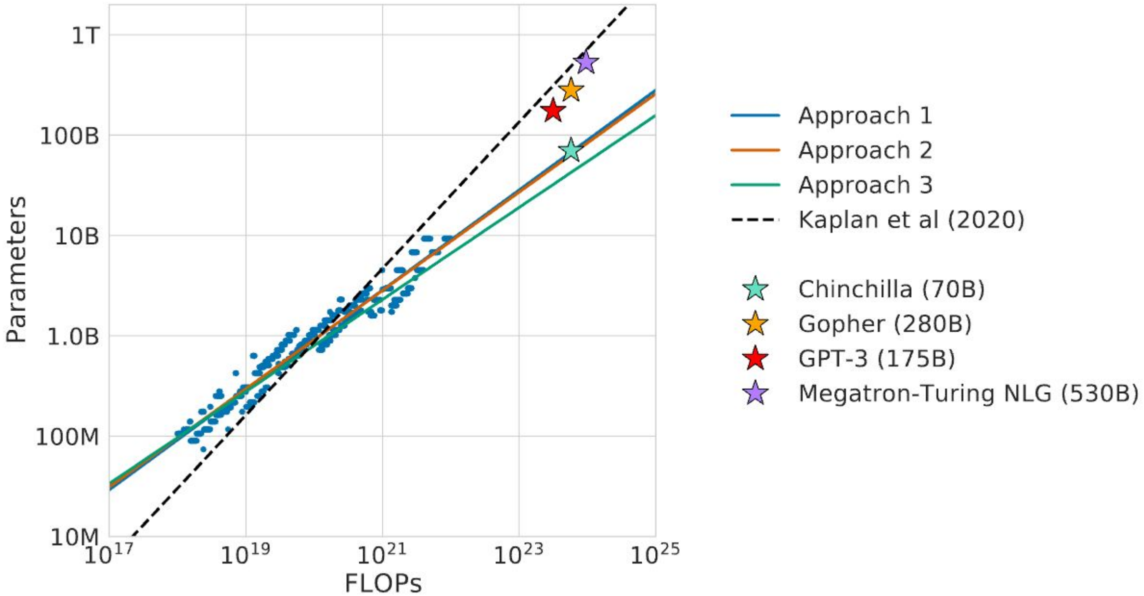


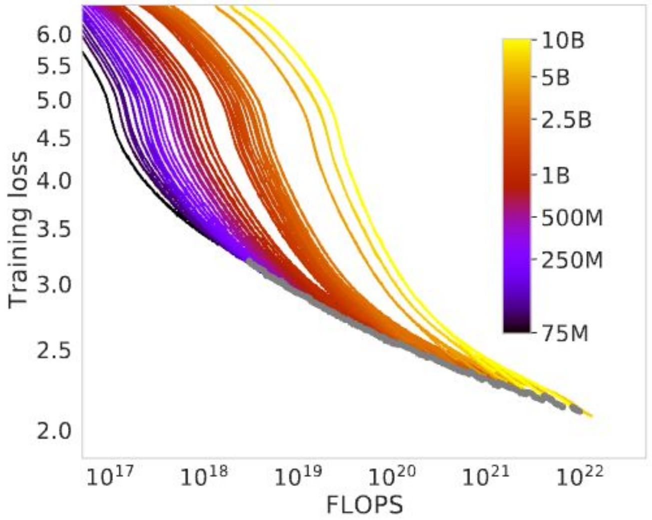
Privacy Risks
Duplicated data is more likely to be memorized and generated [1,2]

LLM "Learning" vs. Memorization
Duplicates between train and test can cause overestimation of perf. [3]

RQ: What are the consequences of dataset de-duplication?

To consider: scaling laws are painful (spooky)!



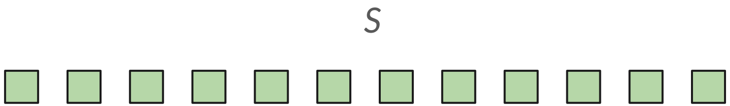


from Hoffman et al.

Deduplication Approaches

<u>k-Substring Matching</u>
Remove verbatim duplicate substrings.

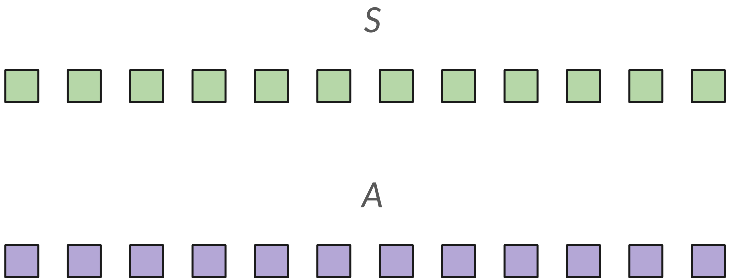
<u>MinHash Matching</u>
Remove full examples with high *n*-gram overlap.



Deduplication Approaches

k-Substring Matching
Remove verbatim duplicate substrings.

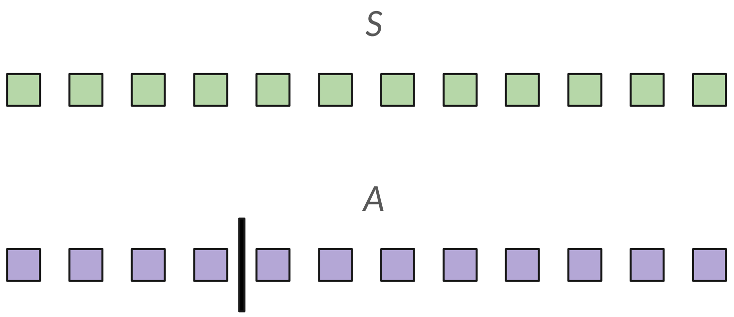
MinHash Matching
Remove full examples with high *n*-gram overlap.



Deduplication Approaches

k-Substring Matching
Remove verbatim duplicate substrings.

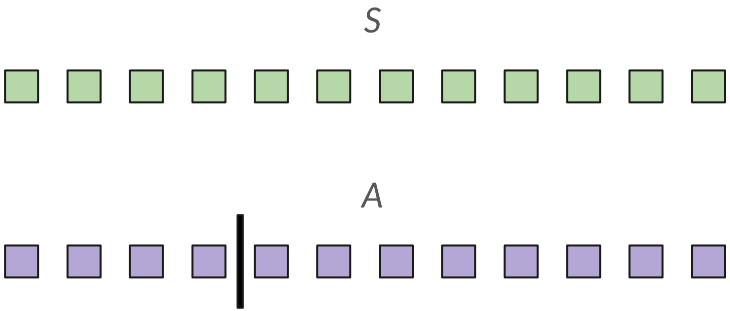
MinHash Matching
Remove full examples with high *n*-gram overlap.



Deduplication Approaches

<u>k-Substring Matching</u>
Remove verbatim duplicate substrings.

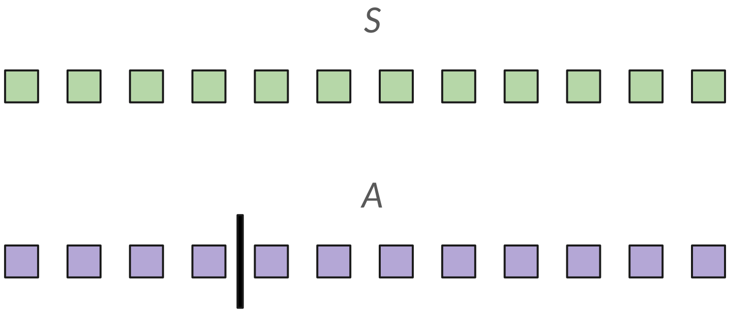
<u>MinHash Matching</u>
Remove full examples with high *n*-gram overlap.



Deduplication Approaches

k-Substring Matching
Remove verbatim duplicate substrings.

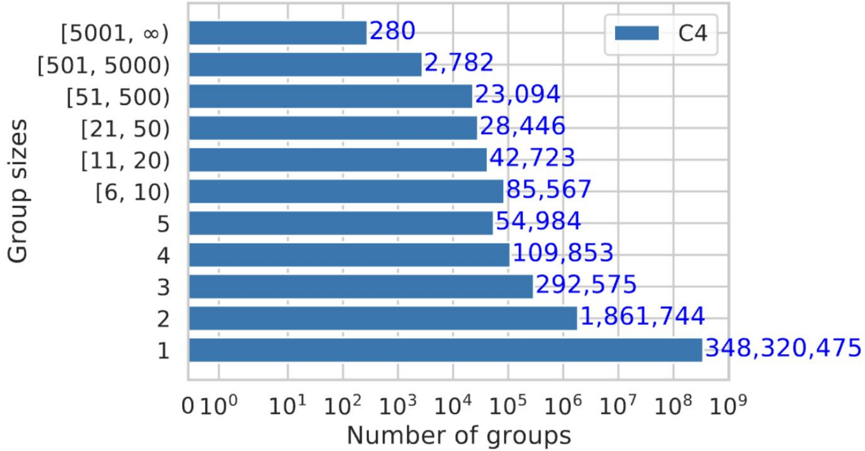
MinHash Matching
Remove full examples with high *n*-gram overlap.



$$\mathrm{Jaccard}(d_i, d_j) = {|d_i \cap d_j|}/{|d_i \cup d_j|}$$

$$\mathrm{EditSim}(x_i, x_j) = 1 - \frac{\mathrm{EditDistance}(x_i, x_j)}{\max(|x_i|, |x_j|)}$$

Results: Dataset Contents



Wide distribution of duplicates in C4, some repeated many times (MinHash).

Removing duplicates would reduce the size of C4 by roughly 3%.

Results: Impact on Trained Models

Model trained on C4 without HashMin duplicates

Model trained on C4 without ExactSubstring duplicates

Model trained on C4

Results: Impact on Trained Models

Model trained on C4 without HashMin duplicates

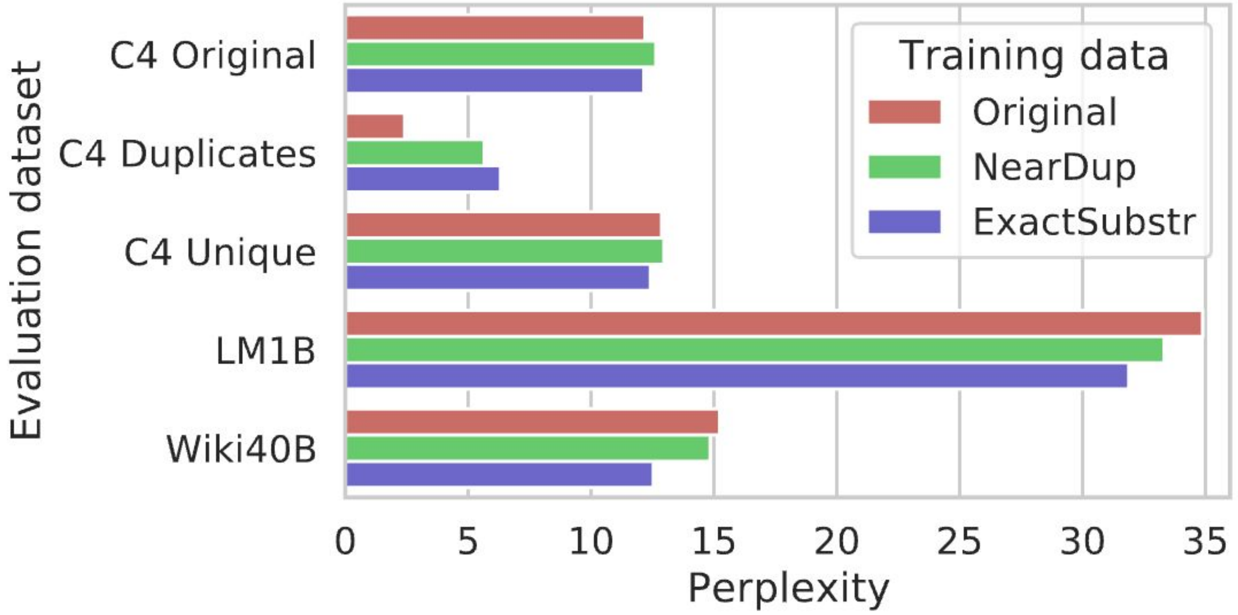
Model trained on C4 without ExactSubstring duplicates

Model trained on C4

Evaluation of perplexity on duplicate or unique examples

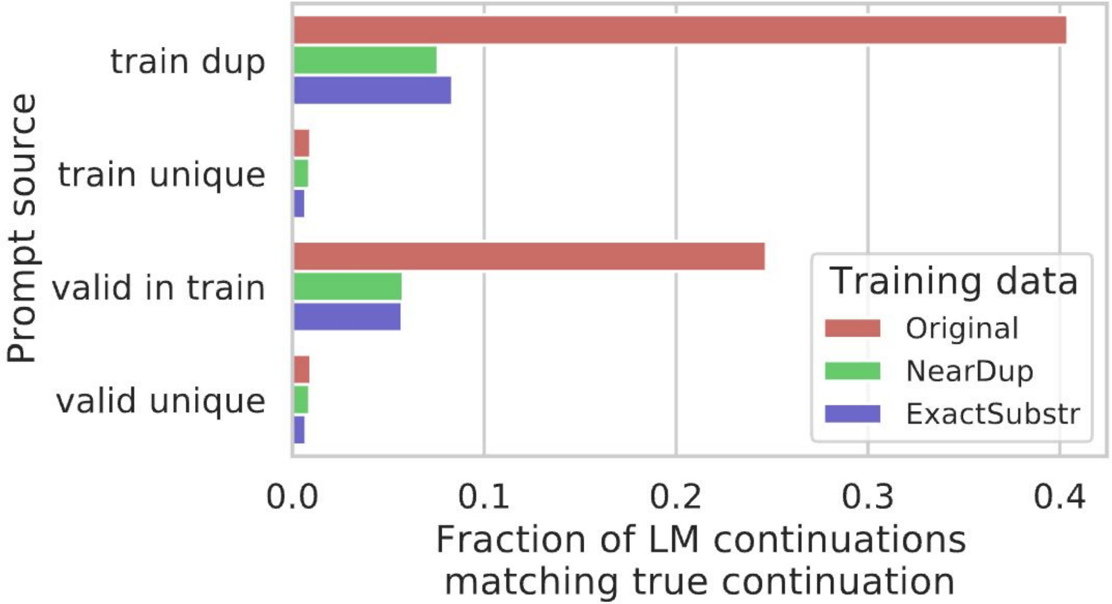
Unique generations of models trained on (de)duplicated data

Results: Impact on Trained Models



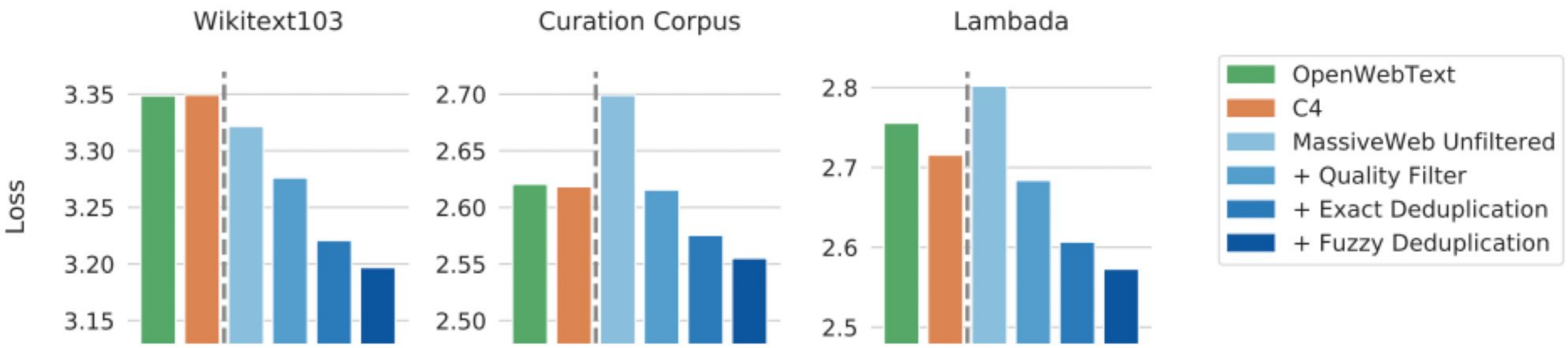
Results: Impact on Trained Models

“The rock parrot (Neophema petrophila) is a species of ….”



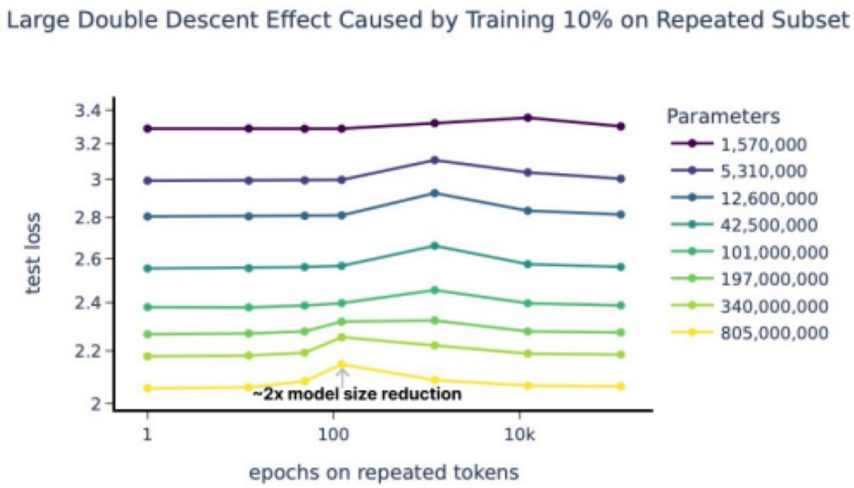
Since then... Gopher

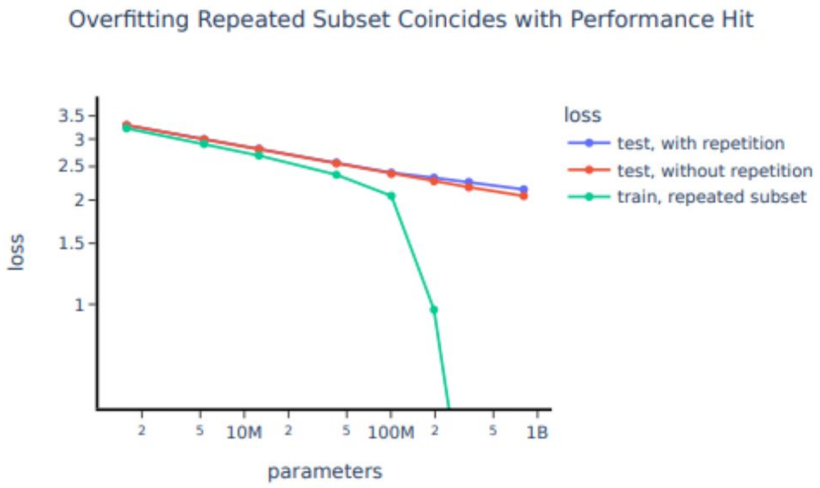
Gopher: data ablations demonstrate that deduplication is helpful on the MassiveText dataset.



Since then... Anthropic

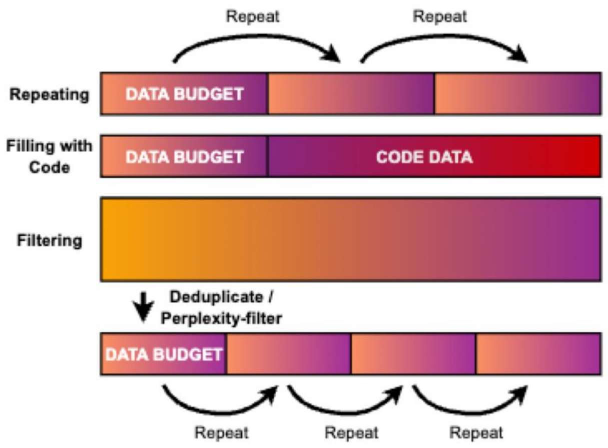
Anthropic: data repetition can cause significant performance degradation.

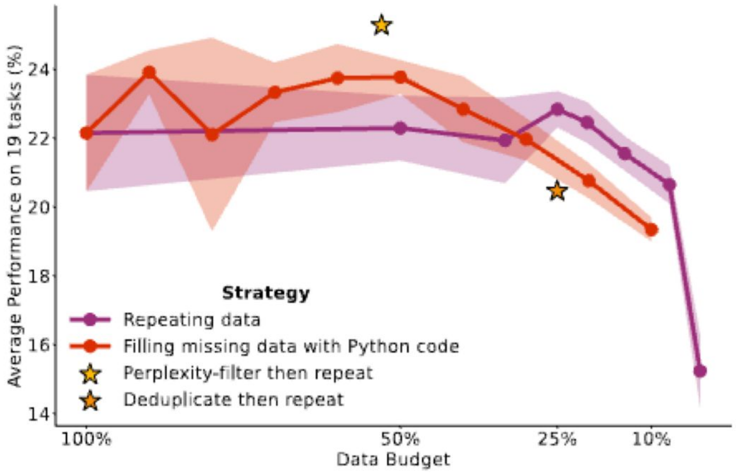




Since then... Datablations

Datablations: data deduplication did not improve downstream task perf.





Many more perspectives for dataset filtering...

Semantic deduplication (in image domain)

Filtering out label errors

(Learning to) filter "low-quality" data

References

1. Allamanis, M., 2019, October. The adverse effects of code duplication in machine learning models of code. In Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (pp. 143-153).
2. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. and Carlini, N., 2021. Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499.
3. Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M. and Gardner, M., 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758.
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9.
5. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. and Oprea, A., 2021. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633-2650).
6. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F. and Zhang, C., 2022. Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646.
7. Magar, I. and Schwartz, R., 2022. Data contamination: From memorization to exploitation. arXiv preprint arXiv:2203.08242.
8. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A. and Hennigan, T., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
9. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S. and Rutherford, E., 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.
10. Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T. and Johnston, S., 2022. Scaling laws and interpretability of learning from repeated data. arXiv preprint arXiv:2205.10487.
11. Muennighoff, N., Rush, A.M., Barak, B., Scao, T.L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T. and Raffel, C., 2023. Scaling Data-Constrained Language Models. arXiv preprint arXiv:2305.16264.

Discussion questions

Language ideologies paper

1. How would you build a “better” filter that is less biased? How should this filter be evaluated?
2. Who should make the judgement call of what constitutes as “high quality” data in large-scale datasets? How does the task matter?
3. The authors recommend “abandoning the notion of a general-purpose corpus.” Does this sound feasible? How does this complicate the notion of “more data is better” for building language models?

Deduplication paper

1. What are some limitations of the deduplication paper? What are some ideas for addressing them?
2. How do the results from deduplication (and filtering more broadly) change your perspective on the scaling law paradigm?
3. What aspects should we consider when we try and define “data quality?”
4. How do we balance LLM memorization of harmful information against innocuous or useful information?

Appendix

Approach 1: Substring matching w/ suffix array

Example: “camel”

Approach 1: Substring matching w/ suffix array

Example: “camel”

0	camel
1	amel
2	mel
3	el
4	l

Approach 1: Substring matching w/ suffix array

Example: "camel"

0	camel
1	amel
2	mel
3	el
4	l

1	amel
0	camel
3	el
4	l
2	mel

Results: Dataset Contents

	% train examples with dup in train	% train examples with dup in valid	% valid with dup in train
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

Table 2: The fraction of examples identified by NEARDUP as near-duplicates.

	% train tokens with dup in train	% train tokens with dup in valid	% valid with dup in train
C4	7.18%	0.75 %	1.38 %
RealNews	19.4 %	2.61 %	3.37 %
LM1B	0.76%	0.016%	0.019%
Wiki40B	2.76%	0.52 %	0.67 %

Table 3: The fraction of tokens (note Table 2 reports the fraction of *examples*) identified by EXACTSUBSTR as part of an exact duplicate 50-token substring.