A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

Shayne Longpre[1], **Gregory Yauney**[2], **Emily Reif**[3], **Katherine Lee**[2,3], **Adam Roberts**[3], **Barret Zoph**[4], **Denny Zhou**[3], **Jason Wei**[4], **Kevin Robinson**[3], **David Mimno**[2], **Daphne Ippolito**[3,5]

[1] MIT [2] Cornell University [3] Google Research [4] OpenAI [5] Carnegie Mellon University

All pretraining data is curated, but data curation decisions are not always disclosed.

1. Practitioners are guided by intuition.
2. Experiments are frequently repeated because results are not disclosed.
3. Data curation has large impact because pretrained models are reused.

We pretrain 28 LMs at the 1.5B-parameter scale on differently-curated pretraining datasets in order to measure the effects of curation choices.

Compute is expensive! But so is dark data & documentation debt.

Datasets: C4 and the Pile

Models: 1.5B-parameter decoder-only autoregressive transformers

Setting: Pretrain, then finetune on downstream tasks individually

Takeaways

1. Stale pretraining data matters and is not overcome by finetuning!
2. Temporal misalignment effects grow with model size.
3. "Quality" filters boost performance, even while reducing training data.
4. Toxicity filters hurt. Inverse toxicity filters can help a lot for some tasks.
5. Data heterogeneity and quantity matter most, especially web and books data.

Data Age

Mismatch in data age between pretraining and evaluation data causes performance degradation.

1. Less impact than finetuning mismatch, but adds up.
2. Release age distributions for pretraining data.

Full Pretraining Dataset

Select Pretraining Data

Scrapes of Common Crawl from different years

2013 2016 2019 2022

Pretrain Models

Evaluation

Datasets with year metadata

- News (PubCLS, NewSum)
- Twitter (PoliAff, TwiERC)
- Science (AIC)

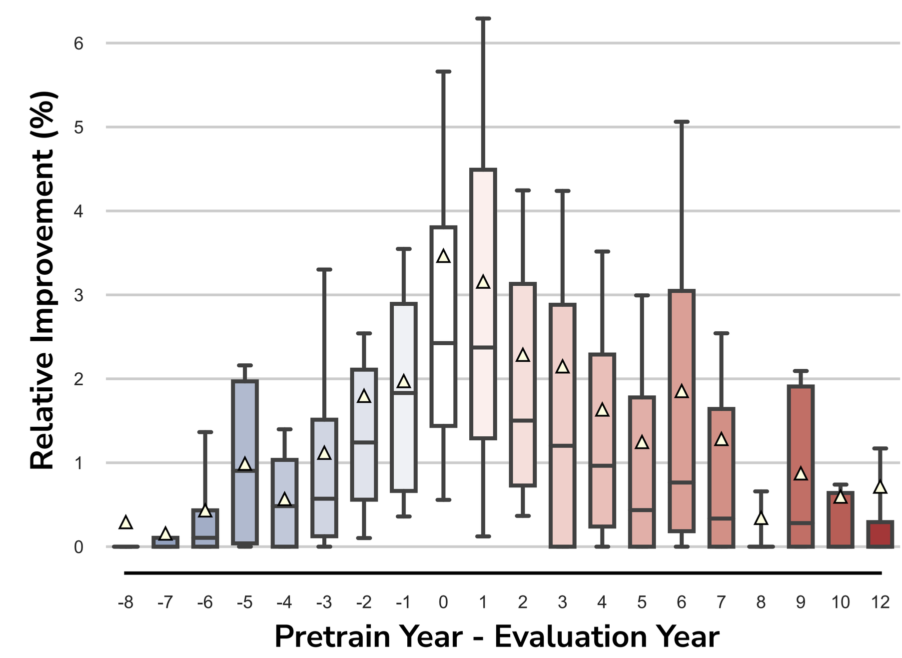
Example dataset: PoliAff

Pretrain Year \ Eval Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
2013	82.7	89.0	91.2	71.2	70.8	74.7	71.5	82.0	82.2	74.9
2016	80.4	88.1	90.6	70.9	72.3	75.8	72.6	82.2	82.4	76.0
2019	80.2	87.8	90.7	70.4	72.0	75.8	73.4	83.1	82.8	75.9
2022	79.4	87.1	89.4	70.8	71.4	75.0	71.0	82.5	83.3	76.8

accuracy after pretraining on 2016 data, evaluating on 2019 data (averaged across finetuning years)

Accuracy is higher when pretraining and eval year are closer in time, even after finetuning

Pretrained models become stale



Temporal degradation happens faster when evaluating old models on new benchmarks.

Toxicity and Quality

Toxicity: Perspective API, a classifier that assigns every document a score from 0 (nontoxic) to 1 (toxic)

Quality: GLaM/PaLM classifier, Wikipedia + books are "high quality", every document gets a score from 0 (high quality) to 1 (low quality)

- Scalable, consistent with current practice
- Many downsides
- Lots of open questions!

Full Pretraining Dataset

Select Pretraining Data

Separate quality & toxicity filtering

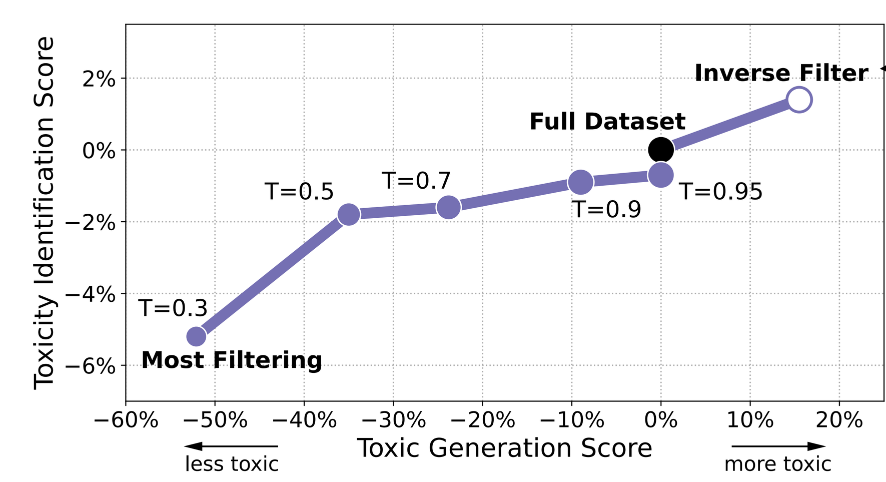
Light … Heavy Inverse

Pretrain Models

Evaluation

1. **Toxicity identification:** Can the model classify text as toxic? (Social Bias Frames, DynaHate, Toxigen)
2. **Toxic generation:** Is generated text considered toxic? (RealToxicityPrompts, RepBias)
3. **Question Answering:** 27 tasks from MRQA & UnifiedQA, categorized by domain

Toxicity filtering induces a tradeoff: reduces toxic generation at the cost of decreased toxicity identification.



If the goal is to identify toxic text, then training on toxic data is more effective

Content filtering impacts downstream QA performance

	Filter	Data	QA domain Wiki	Web	Acad	CS	Mean
Baseline	Full Data	100%	0	0	0	0	0
Toxicity	Light	95%	-2.2	-1.1	0.2	0.2	-0.7
	Heavy	76%	-4.2	-2.4	-1.1	-3.5	-2.7
	Inverse	92%	0.4	-1.4	4.9	2.7	1.7
Quality	Light	91%	1.2	0.7	6.4	6.1	2.5
	Heavy	73%	-0.3	0.8	0.8	6.8	1.2
	Inverse	73%	-5.0	-4.5	-2.7	-6.4	-3.1

1. Toxicity filtering hurts QA performance across domains.
2. Quality filtering improves performance across most domains, despite removing data.

Domain Coverage

Full Pretraining Dataset

Select Pretraining Data

Ablate one domain of the Pile at a time

No Social No Wiki No Books … and so on

Pretrain Models

Evaluation

Question Answering: 27 tasks from MRQA & UnifiedQA, categorized by domain

Heterogeneous domains have biggest effect on QA performance

	Wiki	Web	Biomed	Academic	Common Sense	Contrast Sets	Average
Full Dataset (100%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
No Social (99%)	-0.8	-3.7	0.1	3.5	-3.5	3.5	0.3
No Wiki (98%)	-1.3	-5.3	0.2	0.9	-4.4	7.2	-0.4
No Books (93%)	-3.5	-6.3	0.0	-1.6	-6.5	-4.4	-2.8
No OpenWeb (93%)	-2.0	-4.1	-1.0	0.6	-5.8	-2.9	-1.4
No Legal (91%)	-2.7	-2.9	0.4	0.8	-2.6	-0.4	-0.7
No Academic (87%)	-0.3	-2.5	-0.9	2.2	-1.1	4.3	0.2
No Pubmed (85%)	-0.3	-3.0	-5.8	-1.5	-5.9	3.9	-1.4
No Code (81%)	-0.5	-3.1	-1.2	1.2	-5.8	4.4	-0.2
No CC (73%)	-3.2	-6.2	-4.6	-5.9	-8.0	-5.2	-4.9

1. Removing Books and Common Crawl domains hurt downstream performance most.
2. Targeted data helps for targeted evaluations.

Paper: arxiv.org/pdf/2305.13169.pdf