WHAT'S IN MY AI?

A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher

Alan D. Thompson
LifeArchitect.ai
March 2022
Rev 0

Contents

About the Author

Dr Alan D. Thompson is an AI expert and consultant. With Leta (an AI powered by GPT-3), Alan co-presented a seminar called 'The new irrelevance of intelligence' at the World Gifted Conference in August 2021. His applied AI research and visualizations are featured across major international media, including citations in the University of Oxford's debate on AI Ethics in December 2021. He has held positions as chairman for Mensa International, consultant to GE and Warner Bros, and memberships with the IEEE and IET.

To cite this article:
Thompson, A. D. (2022). *What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher.* https://LifeArchitect.ai/whats-in-my-ai

Abstract

Pre-trained transformer language models have become a stepping stone towards artificial general intelligence (AGI), with some researchers reporting that AGI may evolve[1] from our current language model technology. While these models are trained on increasingly larger datasets, the documentation of basic metrics including dataset size, dataset token count, and specific details of content is lacking. Notwithstanding proposed standards[2] for documentation of dataset composition and collection, nearly all major research labs have fallen behind in disclosing details of datasets used in model training. The research synthesized here covers the period from 2018 to early 2022, and represents a comprehensive view of all datasets—including major components Wikipedia and Common Crawl—of selected language models from GPT-1 to Gopher.

1. Overview

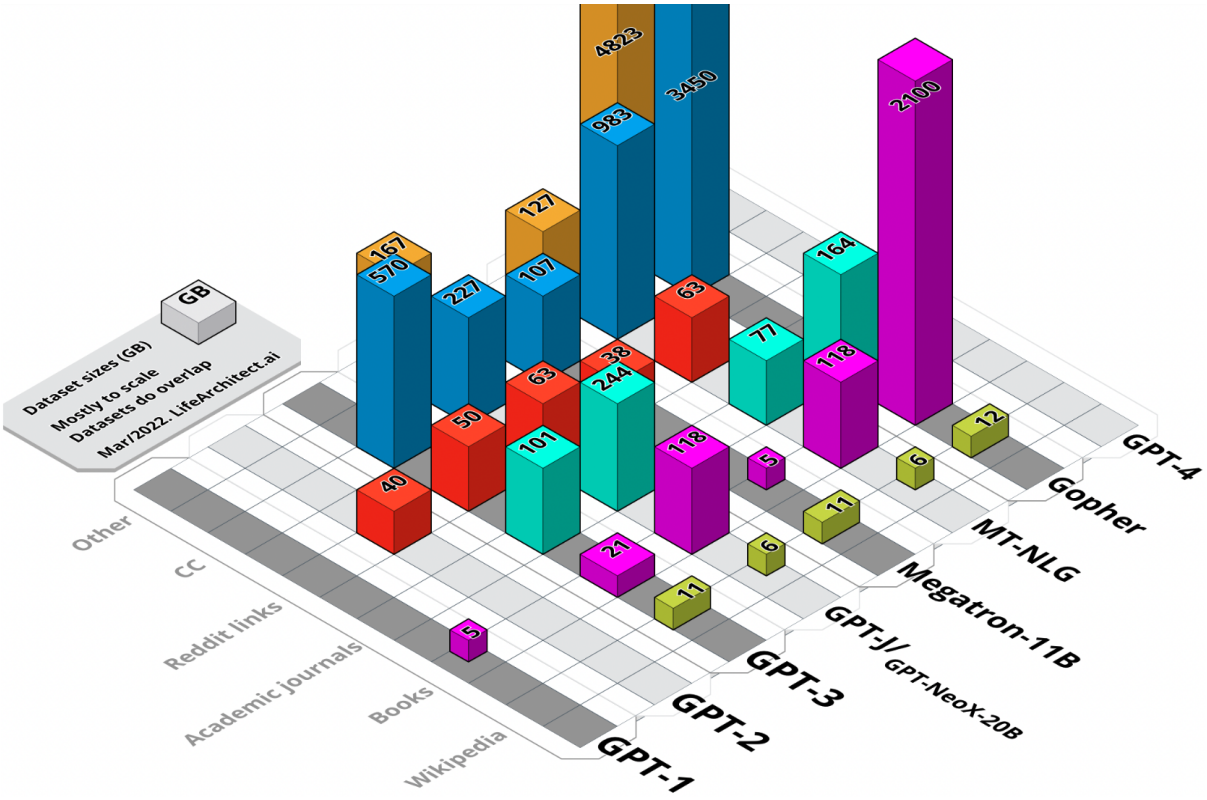

Figure 1. Visual Summary of Major Dataset Sizes. Unweighted sizes, in GB.

Since 2018, large language models have exploded in both development and production use. Major research labs report incredibly high usage by the general public. In March 2021, OpenAI announced[3] that its GPT-3 language model was being used by "more than 300 applications [and generating] an average of 4.5 billion words per day". This is the equivalent of 3.1 million words per minute of new content,

[1] GPT-NeoX-20B paper: pp11, section 6 http://eaidata.bmk.sh/data/GPT_NeoX_20B.pdf
[2] Datasheet for Datasets paper: https://arxiv.org/abs/1803.09010
[3] OpenAI blog: https://openai.com/blog/gpt-3-apps/

generated by just a single model. Notably, these language models are not even completely understood, with Stanford researchers[4] recently admitting that "we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties".

As new AI technology rapidly progresses, there has been a decline in documentation quality about the datasets used to train these models. What is really inside my AI? What is it made of? This article provides a comprehensive synthesis and analysis of datasets used to train modern large language models.

Where primary model papers were opaque, the research synthesized here was collected from secondary and tertiary sources, and often necessitated assumptions to determine final estimates.

In this article, where the primary paper has been clear about a specific detail (for example, token count or dataset size), it is considered 'disclosed', and marked in **bold.**

In many cases, it is necessary to include assumptions to determine final estimates, referencing secondary and tertiary sources where appropriate. In these instances, the detail (for example, token count or dataset size) is considered 'determined', and marked in *italics.*

For each model, datasets are categorized into six groups: Wikipedia, Books, Journals, Reddit links, Common Crawl, and Other.

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		*4.6*					4.6
GPT-2				**40**			40
GPT-3	*11.4*	*21*	*101*	*50*	**570**		753
The Pile v1	**6**	**118**	**244**	**63**	**227**	**167**	825
Megatron-11B	**11.4**	*4.6*		**38**	**107**		161
MT-NLG	*6.4*	*118*	*77*	*63*	*983*	*127*	1374
Gopher	**12.5**	**2100**	*164.4*		*3450*	*4823*	10550

Table 1. Summary of Major Dataset Sizes. Shown in GB. Disclosed in **bold**. Determined in *italics*. Raw training dataset sizes only.

[4] On the Opportunities and Risks of Foundation Models: https://arxiv.org/abs/2108.07258

1.1. Wikipedia
Wikipedia is a free, multilingual, collaborative, online encyclopedia written and maintained by a community of over 300,000 volunteers. As of April 2022, there are over 6.4 million articles in the English Wikipedia, containing over 4 billion words[5]. The text is valuable as it is rigorously referenced, written in expository prose, and spans many languages and domains. Generally, an English-only filtered version of Wikipedia is a popular starting point for use as a dataset by major research labs.

1.2. Books
Narratives consisting of a mix of fiction and nonfiction books are useful for coherent storytelling and responses. Includes datasets like Project Gutenberg and Smashwords (Toronto BookCorpus/BookCorpus).

1.3. Journals
Papers in preprint and published journals provide a solid and rigorous foundation for datasets, as academic writing typically demonstrates methodical, rational, and meticulous output. Includes datasets like ArXiv and The National Institutes of Health (US).

1.4. Reddit links
WebText is a large dataset sourced from a general web scrape of all outbound links from social media platform Reddit, where the links have received at least 3 upvotes. This is used as a heuristic indicator for popular content, perhaps suggesting higher quality links and subsequent text data.

1.5. Common Crawl
Common Crawl is a large dataset of website crawls from 2008-present, including raw web pages, metadata, and text extracts. Common Crawl includes text from diverse languages and domains. An English-only filtered version of Common Crawl called C4 is a popular starting point for use as a dataset by major research labs.

1.6. Other
Datasets which do not fit into any of the groups above include code datasets like GitHub, conversation forums like StackExchange, and video subtitle datasets.

2. Common Datasets

Since 2019, most transformer-based large language models (LLMs) rely on large datasets from English Wikipedia, and from the Common Crawl. In this section, we provide a high-level overview of English Wikipedia by category, and the top domains in Common Crawl using Google C4[6] (Colossal Clean Crawled Corpus) based on the

[5] Size of Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
[6] C4 dataset: https://www.tensorflow.org/datasets/catalog/c4

Common Crawl dataset[7], with reference to the comprehensive analysis conducted by Jesse Dodge and team at AllenAI (AI2)[8].

2.1. Wikipedia (English) Analysis

Detail on Wikipedia by category[9] is included below, with coverage of 1001 random articles sampled in 2015, with researchers noting the stability of spread over time. Assuming an 11.4GB cleaned and filtered version of English Wikipedia with 3 billion tokens, we can determine category sizes and tokens.

Rank	Category	Percentage	Size (GB)	Tokens (M)
1	Biography	**27.8%**	*3.1*	*834*
2	Geography	**17.7%**	*1.9*	*531*
3	Culture and Arts	**15.8%**	*1.7*	*474*
4	History	**9.9%**	*1.1*	*297*
5	Biology, Health, and Medicine	**7.8%**	*0.9*	*234*
6	Sports	**6.5%**	*0.7*	*195*
7	Business	**4.8%**	*0.5*	*144*
8	Other society	**4.4%**	*0.5*	*132*
9	Science & Math	**3.5%**	*0.4*	*105*
10	Education	**1.8%**	*0.2*	*54*
	Total	**100%**	*11.4*	*3000*

Table 2. English Wikipedia Dataset Categories. Disclosed in **bold**. Determined in *italics*.

2.2. Common Crawl Analysis

Based on work in the C4 paper by AllenAI (AI2), we can determine both token count and overall percentage of each domain for the filtered English C4 dataset, which is 305GB with 156B tokens.

Rank	Domain	Tokens (M)	%	Rank	Domain	Tokens (M)	%
1	**Google Patents**	*750*	*0.48%*	13	**Frontiers Media**	*60*	*0.04%*
2	**The NY Times**	*100*	*0.06%*	14	**Business Insider**	*60*	*0.04%*

[7] Common Crawl website: https://commoncrawl.org/

[8] C4 paper: https://arxiv.org/abs/2104.08758 pp2, Figure 1 right

[9] Wikipedia categories: https://en.wikipedia.org/wiki/User:Smallbones/1000_random_results: "What topics does Wikipedia cover? Has the coverage changed over time? These and similar questions are examined using 1001 random articles sampled in December, 2015... These proportions are fairly stable over time... Biography (27.8%) Geography (17.7%) Culture and arts (15.8%) History (9.9%) Biology, health, and medicine (7.8%) Sports (6.5%) Business (4.8%) Other society (4.4%) Science & Math (3.5%) Education (1.8%)"

Rank	Domain	Tokens (M)	%	Rank	Domain	Tokens (M)	%
3	**Los Angeles Times**	*90*	*0.06%*	15	**Chicago Tribune**	*59*	*0.04%*
4	**The Guardian**	*90*	*0.06%*	16	**Booking.com**	*58*	*0.04%*
5	**PLoS**	*90*	*0.06%*	17	**The Atlantic**	*57*	*0.04%*
6	**Forbes**	*80*	*0.05%*	18	**Springer Link**	*56*	*0.04%*
7	**HuffingtonPost**	*75*	*0.05%*	19	**Al Jazeera**	*55*	*0.04%*
8	**Patents.com**	*71*	*0.05%*	20	**Kickstarter**	*54*	*0.03%*
9	**Scribd**	*70*	*0.04%*	21	**FindLaw Caselaw**	*53*	*0.03%*
10	**Washington Post**	*65*	*0.04%*	22	**NCBI**	*53*	*0.03%*
11	**The Motley Fool**	*61*	*0.04%*	23	**NPR**	*52*	*0.03%*
12	**IPFS**	*60*	*0.04%*				
					Total	*2.2B*	*1.42%*
					Remainder	*153.8B*	*98.58%*

Table 3. C4: Top 23 Domains (excluding Wikipedia). Disclosed in **bold**. Determined in *italics*.

3. GPT-1 Dataset

The GPT-1 model was released by OpenAI in 2018, with 117M parameters. The paper was unclear about the source and contents of the training dataset used[10]. The paper misspelled 'BookCorpus' as 'BooksCorpus'. BookCorpus is based on free books written by unpublished authors and sourced from Smashwords, an ebook website that describes itself as "the world's largest distributor of indie ebooks". The dataset has also been called the Toronto BookCorpus. Several recreations of the BookCorpus determined the size of that dataset to be 4.6GB[11].

In 2021, a comprehensive retrospective analysis was conducted on the BookCorpus dataset[12]. Corrections were provided for genre grouping and book counts by genre. Further detail on book genres inside the dataset includes:

Count	Genre	Book count	Percentage (Book count / 11038)
1	Romance	**2880**	**26.1%**
2	Fantasy	**1502**	**13.6%**
3	Science Fiction	**823**	**7.5%**
4	New Adult	**766**	**6.9%**
5	Young Adult	**748**	**6.8%**
6	Thriller	**646**	**5.9%**
7	Mystery	**621**	**5.6%**
8	Vampires	**600**	**5.4%**

[10] GPT-1 paper: pp4 "We use the BooksCorpus dataset for training the language model."
[11] https://huggingface.co/datasets/bookcorpus: "Size of the generated dataset: 4629.00 MB"
[12] BookCorpus Retrospective Datasheet paper: pp9 https://arxiv.org/abs/2105.05241

Count	Genre	Book count	Percentage (Book count / 11038)
9	Horror	**448**	**4.1%**
10	Teen	**430**	**3.9%**
11	Adventure	**390**	**3.5%**
12	Other	**360**	**3.3%**
13	Literature	**330**	**3.0%**
14	Humor	**265**	**2.4%**
15	Historical	**178**	**1.6%**
16	Themes	**51**	**0.5%**
	Total	11038	100%

Table 4. BookCorpus Genres. Disclosed in **bold**. Determined in *italics*.

In subsequent dataset recreations (i.e. for The Pile v1 and others), further filtering was applied to: exclude the 'Vampires' genre, decrease the percentage of books in the 'Romance' genre, increase the percentage of books in the 'Historical' genre, substantially increase the number of books collected.

3.1. GPT-1 Dataset Summary

The final dataset summary analysis of the GPT-1 model is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB		*4.6*					4.6
Tokens		*1.3*					1.3

Table 5. GPT-1 Datasets Summary. Shown in GB. Disclosed in **bold**. Determined in *italics*.

4. GPT-2 Dataset

The GPT-2 model was released by OpenAI in 2019, with 1.5B parameters. The GPT-2 paper was clear about the size[13] of the training dataset used, but not about the contents. The GPT-2 model card[14] (in the GPT-2 GitHub repository) was clear about the model contents.

Token count can be derived from the GPT-3 paper, which uses an extended version of WebText for 19B tokens. It has been assumed that this 2020 extended version has 12 months of additional data, and so it is possible that it is around 25% larger than the 2019 GPT-2 version[15]. The GPT-2 final token count is determined to be around 15B.

Detail on the contents of WebText as a percentage of the dataset can be determined assuming that the model card is showing count of links, each of which can be divided by the total of 45 million links, as noted in the GPT-2 paper.

The determined token count of 15B can then be used to find a token count per domain. Note that of the Top 1,000 domains available, only the Top 50 domains are shown here.

Rank	Domain	Links (M)	%	Tokens (M)	Rank	Domain	Links (M)	%	Tokens (M)
1	**Google**	**1.54**	*3.4%*	*514*	**26**	**Independent**	**0.11**	*0.2%*	*35*
2	**Archive**	**0.60**	*1.3%*	*199*	**27**	**Etsy**	**0.11**	*0.2%*	*35*
3	**Blogspot**	**0.46**	*1.0%*	*152*	**28**	**Craigslist**	**0.10**	*0.2%*	*33*
4	**GitHub**	**0.41**	*0.9%*	*138*	**29**	**BusinessInsider**	**0.09**	*0.2%*	*31*
5	**The NY Times**	**0.33**	*0.7%*	*111*	**30**	**Telegraph**	**0.09**	*0.2%*	*31*
6	**WordPress**	**0.32**	*0.7%*	*107*	**31**	**Wizards**	**0.09**	*0.2%*	*30*
7	**WashingtonPost**	**0.32**	*0.7%*	*105*	**32**	**USAtoday**	**0.08**	*0.2%*	*28*
8	**Wikia**	**0.31**	*0.7%*	*104*	**33**	**TheHill**	**0.08**	*0.2%*	*27*
9	**BBC**	**0.31**	*0.7%*	*104*	**34**	**NHL**	**0.08**	*0.2%*	*27*
10	**TheGuardian**	**0.25**	*0.5%*	*82*	**35**	**FoxNews**	**0.08**	*0.2%*	*26*
11	**eBay**	**0.21**	*0.5%*	*70*	**36**	**Taobao**	**0.08**	*0.2%*	*26*
12	**Pastebin**	**0.21**	*0.5%*	*70*	**37**	**Bloomberg**	**0.08**	*0.2%*	*26*
13	**CNN**	**0.20**	*0.4%*	*66*	**38**	**NPR**	**0.08**	*0.2%*	*26*
14	**Yahoo**	**0.20**	*0.4%*	*65*	**39**	**MLB**	**0.08**	*0.2%*	*26*
15	**HuffingtonPost**	**0.19**	*0.4%*	*62*	**40**	**LA Times**	**0.08**	*0.2%*	*26*

[13] GPT-2 paper: pp3 “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....WebText, contains the text subset of these 45 million links... which does not include links created after Dec 2017 and which after de-duplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text. We removed all Wikipedia documents from WebText...”

[14] GPT-2 model card: https://github.com/openai/gpt-2/blob/master/model_card.md: “We’ve published a list of the top 1,000 domains present in WebText and their frequency. The top 15 domains by volume in WebText are: Google, Archive, Blogspot, GitHub, NYTimes, Wordpress, Washington Post, Wikia, BBC, The Guardian, eBay, Pastebin, CNN, Yahoo!, and the Huffington Post.”

[15] GPT-3 paper: “WebText2: 19 billion tokens. [Alan: WebText2 is slightly extended from WebText, so we can subtract 20% to get 15 billion]”

Rank	Domain	Links (M)	%	Tokens (M)
16	**Go**	**0.19**	*0.4%*	*62*
17	**Reuters**	**0.18**	*0.4%*	*61*
18	**IMDb**	**0.18**	*0.4%*	*61*
19	**Goo**	**0.16**	*0.4%*	*54*
20	**NIH**	**0.14**	*0.3%*	*47*
21	**CBC**	**0.14**	*0.3%*	*45*
22	**Apple**	**0.13**	*0.3%*	*43*
23	**Medium**	**0.13**	*0.3%*	*42*
24	**DailyMail**	**0.12**	*0.3%*	*40*
25	**SteamPowered**	**0.11**	*0.2%*	*36*

Rank	Domain	Links (M)	%	Tokens (M)
41	**Megalodon**	**0.08**	*0.2%*	*25*
42	**ESPN**	**0.07**	*0.2%*	*24*
43	**KickStarter**	**0.07**	*0.2%*	*24*
44	**BreitBart**	**0.07**	*0.2%*	*24*
45	**ABC**	**0.07**	*0.2%*	*23*
46	**NewEgg**	**0.07**	*0.2%*	*23*
47	**WWE**	**0.07**	*0.1%*	*22*
48	**MyAnimeList**	**0.07**	*0.1%*	*22*
49	**Microsoft**	**0.07**	*0.1%*	*22*
50	**Buzzfeed**	**0.06**	*0.1%*	*22*
	Total	*9.3M*	*20.7%*	
	Remainder	*35.7*	*79.3%*	

Table 6. WebText: Top 50 Domains. Disclosed in **bold**. Determined in *italics*.

4.1. GPT-2 Dataset Summary

The final dataset summary analysis of the GPT-2 model is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB				**40**			40
Tokens				*15*			15

Table 7. GPT-2 Datasets Summary. Disclosed in **bold**. Determined in *italics*.

5. GPT-3 Datasets

The GPT-3 model was released by OpenAI in 2020, with 175B parameters. The paper was clear about the token counts[16] of the training datasets used, but the contents and sizes were unclear (except for the size of Common Crawl[17]).

Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Common Crawl (filtered)	**410B**	-	*0.71*	*1:1.9*	**570**
WebText2	**19B**	*25% > WebText*	*0.38*	*1:2.6*	*50*
Books1	**12B**	*Gutenberg*	*0.57*	*1:1.75*	*21*
Books2	**55B**	*Bibliotik*	*0.54*	*1:1.84*	*101*
Wikipedia	**3B**	*See RoBERTa*	*0.26*	*1:3.8*	*11.4*
Total	**499B**				*753.4GB*

Table 8. GPT-3 Datasets. Disclosed in **bold**. Determined in *italics*.

5.1. GPT-3: Concerns with Dataset Analysis of Books1 and Books2

Of particular concern, sizes and sources for datasets Books1 (12B tokens) and Books2 (55B tokens) were not disclosed by OpenAI in the GPT-3 paper. Several hypotheses have been put forward about the sources of these two datasets, including similar datasets from LibGen[18] and Sci-Hub, both of which are too large to match, having datasets measured in many Terabytes.

5.2. GPT-3: Books1

The GPT-3 Books1 dataset cannot be the same as the GPT-1 BookCorpus, due to the Books1 dataset's noted larger size of 12 billion tokens. The GPT-1 BookCorpus was described in a cited paper[19] as having 984.8M words, which may be equivalent to only 1.3B tokens (984.8 words x 1.3 word to token multiplier).

---

[16] GPT-2 paper: pp3 "GPT-3: pp9, Table 2.2 "CC: 410B tokens. WebText2: 19B tokens. Books1: 12B tokens. Books2: 55B tokens. Wiki: 3B tokens"

[17] GPT-3 paper: pp8 "we added several curated high-quality datasets, including an expanded version of the WebText dataset, collected by scraping links over a longer period of time, and first described in, two internet-based books corpora (Books1 and Books2) and English-language Wikipedia... The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens"

[18] BookCorpus repo: https://github.com/soskek/bookcorpus/issues/27: "books3.tar.gz seems to be similar to OpenAI's mysterious "books2" dataset referenced in their papers. Unfortunately OpenAI will not give details, so we know very little about any differences. People suspect it's "all of libgen", but it's purely conjecture. Nonetheless, books3 is "all of bibliotik"..."

[19] BookCorpus paper: https://arxiv.org/abs/1506.06724: "# of words: 984,846,357 [Alan: BookCorpus is then 1.3B tokens. We want 12-55B tokens]"

It is possible that Books1 aligns with Gutenberg via the Standardized Project Gutenberg Corpus (SPGC), an open science approach to a curated version of the complete PG data of the Gutenberg Project. The SPGC is 12 billion tokens[20], and about 21GB[21].

5.3. GPT-3: Books2

It is possible that Books2 (55B tokens) aligns with Bibliotik, and a dataset made up of data from this source was collected by EleutherAI as part of The Pile v1. That version of Bibliotik is 100.96GB[22], which gives a determined count of only 25B tokens; lower than the 55B tokens disclosed for Books2. However, using the 'tokens per byte' ratio of SPGC (around 1:1.75), the Bibliotik token count and size would more closely match that of Books2.

5.4. GPT-3 Dataset Summary

A list of top resources for datasets using Wikipedia + CommonCrawl + WebText is outlined in Appendix A. The final dataset summary analysis of the GPT-3 model is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	*11.4*	*21*	*101*	*50*	**570**		753
Tokens	**3**	**12**	**55**	**19**	**410**		499

Table 9. GPT-3 Datasets Summary. Disclosed in **bold**. Determined in *italics*.

6. The Pile v1 (GPT-J & GPT-NeoX-20B) datasets

The Pile v1 dataset was released by EleutherAI in 2021, and the dataset has been used to train many models including GPT-J, GPT-NeoX-20B, and as a partial dataset for other models including MT-NLG. The Pile v1 paper was very clear about the sources and sizes of the training datasets used. With the addition of token counts, The Pile v1 paper should be used as the gold standard for future documentation of datasets.

Further detail on the token counts can be determined using information made available in the paper, Tables 1 (size in GB) and 7 (Tokens per byte)[23].

[20] Gutenberg paper: https://arxiv.org/abs/1812.08092: "we present the Standardized Project Gutenberg Corpus (SPGC), an open science approach to a curated version of the complete PG data containing more than 50,000 books and more than 3×109 word-tokens [Alan: equivalent to about 12B BPE tokens, see below]"

[21] Gutenberg repo: https://zenodo.org/record/2422561 "uncompressed size: 3GB (counts) + 18GB (tokens) [21GB total]"

[22] The Pile v1 paper: "Books3 (Bibliotik tracker): 100.96GB" [Alan: multiplied by Tokens per byte of 0.2477 = 25B tokens]

[23] The Pile v1 paper: pp3, Table 1 for datasets. pp28, Table 7 for Tokens per byte.

Count	Dataset	Raw Size (GB)	Tokens per byte (Tokens / bytes)	Tokens (Raw Size x TpB)
1	Common Crawl (Pile-CC)	**227.12**	**0.2291**	*52.0B*
2	PubMed Central	**90.27**	**0.3103**	*28.0B*
3	Books3	**100.96**	**0.2477**	*25.0B*
4	OpenWebText2	**62.77**	**0.2434**	*15.3B*
5	ArXiv	**56.21**	**0.3532**	*19.9B*
6	Github	**95.16**	**0.4412**	*42.0B*
7	FreeLaw	**51.15**	**0.2622**	*13.4B*
8	Stack Exchange	**32.20**	**0.3436**	*11.1B*
9	USPTO Background	**22.90**	**0.2116**	*4.8B*
10	PubMed Abstracts	**19.26**	**0.2183**	*4.2B*
11	Gutenberg	**10.88**	**0.2677**	*2.9B*
12	OpenSubtitles	**12.98**	**0.2765**	*3.6B*
13	Wikipedia	**6.38**	**0.2373**	*1.5B*
14	DM Mathematics	**7.75**	**0.8137**	*6.3B*
15	Ubuntu IRC	**5.52**	**0.3651**	*2.0B*
16	BookCorpus2	**6.30**	**0.2430**	*1.5B*
17	EuroParl	**4.59**	**0.3879**	*1.8B*
18	HackerNews	**3.90**	**0.2627**	*1.0B*
19	YouTubeSubtitles	**3.73**	**0.4349**	*1.6B*
20	PhilPapers	**2.38**	**0.2688**	*0.6B*
21	NIH ExPorter	**1.89**	**0.1987**	*0.4B*
22	Enron Emails	**0.88**	**0.3103**	*0.3B*
	Total	**825.18GB**		*239.2B*

Table 10. The Pile v1 Dataset. Disclosed in **bold**. Determined in *italics*.

6.1. The Pile v1 Grouped Datasets

To determine dataset sizes for categories like 'Books', 'Journals', and 'CC', datasets have been grouped according to component, as shown in the table below.

Count	Dataset	Raw Size (GB)	Tokens per byte (Tokens / bytes)	Tokens (Raw Size x TpB)
	Books			
1	Books3	**100.96**	**0.3103**	*28.0B*
2	Gutenberg	**10.88**	**0.2677**	*2.9B*
3	BookCorpus2	**6.30**	**0.2430**	*1.5B*
	Books total	*118.14*		*32.4*
	Journals			
4	PubMed Central	**90.27**	**0.3103**	*28.0B*

Alan D. Thompson. 2022. *What's in my AI? A Comprehensive Analysis of Datasets used to Train GPT...*
https://LifeArchitect.ai/whats-in-my-ai

Count	Dataset	Raw Size (GB)	Tokens per byte (Tokens / bytes)	Tokens (Raw Size x TpB)
5	ArXiv	**56.21**	**0.3532**	*19.9B*
6	FreeLaw	**51.15**	**0.2622**	*13.4B*
7	USPTO Background	**22.90**	**0.3436**	*11.1B*
8	PubMed Abstracts	**19.26**	**0.2183**	*4.2B*
9	PhilPapers	**2.38**	**0.2688**	*0.6B*
10	NIH ExPorter	**1.89**	**0.1987**	*0.4B*
	Journals total	*244.06*		*77.6B*
	Other			
11	Github	**95.16**	**0.4412**	*42.0B*
12	Stack Exchange	**32.20**	**0.3436**	*11.1B*
13	OpenSubtitles	**12.98**	**0.2765**	*3.6B*
14	DM Mathematics	**7.75**	**0.8137**	*6.3B*
15	Ubuntu IRC	**5.52**	**0.3651**	*2.0B*
16	EuroParl	**4.59**	**0.3879**	*1.8B*
17	HackerNews	**3.90**	**0.2627**	*1.0B*
18	YouTubeSubtitles	**3.73**	**0.4349**	*1.6B*
19	Enron Emails	**0.88**	**0.3103**	*0.3B*
	Other total	*166.71*		*69.7*

Table 11. The Pile v1 Grouped Datasets (excluding Wikipedia, CC, and WebText). Disclosed in **bold**. Determined in *italics*.

6.2. The Pile v1 Dataset Summary

The final dataset summary analysis of the Pile v1 dataset and GPT-J and GPT-NeoX-20B models is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	**6**	**118**	**244**	**63**	**227**	**167**	825
Tokens	*1.4*	*32*	*77*	*15*	*52*	*70*	247

Table 12. The Pile v1 Datasets Summary. Disclosed in **bold**. Determined in *italics*.

7. Megatron-11B & RoBERTa Datasets

The RoBERTa model was released by Meta AI (then Facebook AI) and the University of Washington in 2019, with 125M parameters. The Megatron-11B model was released by Meta AI (then Facebook AI) in 2020, with 11B parameters. It used the same training dataset as RoBERTa. The RoBERTa[24] paper was clear about the contents of the training datasets used, though cited papers (BERT[25] and Stories[26]) had to be referenced to determine final sizes.

BookCorpus: determined to be 4.6GB, referenced as in the GPT-1 section above.

Wikipedia: disclosed as "BookCorpus plus English Wikipedia… 16GB." Wikipedia was determined to be 11.4GB after subtracting the BookCorpus (4.6GB, as referenced in the GPT-1 section above).

CC-News: disclosed as 76GB after filtering.

OpenWebText: disclosed as 38GB.

Stories: disclosed as 31GB. Note that this dataset is Common Crawl content "based on questions in commonsense reasoning tasks", and does not fit into the 'Books' category in this article. Instead, it is combined with the CC-News dataset (76GB), making a Common Crawl total of 107GB.

7.1. Megatron-11B & RoBERTa Dataset Summary

The final dataset summary analysis of the Megatron-11B and RoBERTa models is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	**11.4**	*4.6*		**38**	**107**		161
Tokens	*3*	*1*		*15*	*143*		162

Table 13. Megatron-11B & RoBERTa Datasets Summary. Disclosed in **bold**. Determined in *italics*.

[24] RoBERTa paper: https://arxiv.org/abs/1907.11692 "BOOKCORPUS plus English WIKIPEDIA. This is the original data used to train BERT. (16GB)."
[25] BERT paper: https://arxiv.org/abs/1810.04805 "BERT is trained on the BooksCorpus (800M words) and Wikipedia (2,500M words)."
[26] Stories paper: https://arxiv.org/abs/1806.02847 pp5-6: "Namely, we build a customized text corpus based on questions in commonsense reasoning tasks. It is important to note that this does not include the answers and therefore does not provide supervision to our resolvers. In particular, we aggregate documents from the CommonCrawl dataset that has the most overlapping n-grams with the questions… Documents in this corpus contain long series of events with complex references from several pronouns. The top 0.1% of highest ranked documents is chosen as our new training corpus. We name this dataset STORIES since most of the constituent documents take the form of a story with long chain of coherent events."

8. MT-NLG Datasets

The MT-NLG model was released by NVIDIA and Microsoft in 2021, with 530B parameters. MT-NLG is the successor to Microsoft Turing NLG 17B and NVIDIA Megatron-LM 8.3B. The MT-NLG paper was very clear about the sources and tokens of the training datasets used, though sizes were not explicitly noted.

Further detail on sizes can be determined using information made available in The Pile v1 paper, as covered earlier in this article. Despite using the same components, it should be noted that reported component sizes in MT-NLG and The Pile v1 are different, due to different filtering and deduplication methods employed by researchers from Eleuther AI (The Pile v1 dataset) and Microsoft/NVIDIA (MT-NLG model).

8.1. Common Crawl in MT-NLG

Pile-CC: disclosed as 49.8B tokens, determined to be around 227.12GB, referenced in The Pile v1 section above.

CC-2020-50: disclosed as 68.7B tokens, assume a tokens per byte rate of 0.25 TpB = 274.8GB.

CC-2021-04: disclosed as 82.6B tokens, assume a tokens per byte rate of 0.25 TpB = 330.4GB

RealNews (from RoBERTa/Megatron-11B): disclosed as 21.9B tokens. Using the RealNews paper[27], the dataset is determined to be 120GB.

CC-Stories (from RoBERTa/Megatron-11B): disclosed as 5.3B tokens, determined to be 31GB as in RoBERTa section above.

Total Common Crawl data from all sources above is determined to be 983.32GB, with 228.3B tokens.

[27] RealNews paper: https://arxiv.org/abs/1905.12616v3 “After deduplication, RealNews is 120 gigabytes without compression.”

8.2. MT-NLG Grouped Datasets

Count	Dataset	Size from Pile v1 (GB)	Pile Tokens (see above)	▲ ▼	MT-NLG Tokens
1	Common Crawl (Pile-CC)	*227.12*	52.0B	▼	**49.8B**
2	Other CC as above	*756.2*	-		**178.5B**
	Common Crawl total	*983.3*			*228.3B*
3	OpenWebText2	*62.77*	15.3B	▼	**14.8B**
4	Wikipedia	*6.38*	1.5B	▲	**4.2B**
	Books				
5	Books3	*100.96*	25.0B	▲	**25.7B**
6	BookCorpus2	*6.30*	1.5B	-	**1.5B**
7	Gutenberg	*10.88*	2.9B	▼	**2.7B**
	Books total	*118.14*	*29.4*		*29.9*
	Journals				
8	PubMed Abstracts	*19.26*	4.2B	▲	**4.4B**
9	NIH ExPorter	*1.89*	0.4B	▼	**0.3B**
10	ArXiv	*56.21*	19.9B	▲	**20.8B**
	Journals total	*77.36*	*24.5*	▲	*25.5*
	Other				
11	Stack Exchange	*32.20*	11.1B	▲	**11.6B**
12	GitHub	*95.16*	42.0B	▼	**24.3B**
	Other total	*127.36*	*53.1B*		*35.9*
	Total	*1375.31*			*338.6B*

Table 14. MT-NLG Grouped Datasets. Disclosed in **bold**. Determined in *italics*.

8.3. MT-NLG Dataset Summary

The final dataset summary analysis of the MT-NLG model is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	*6.4*	*118.1*	*77.4*	*62.8*	*983.3*	*127.3*	1375
Tokens	**4.2**	*29.9*	*228.3*	**14.8**	*143*	*35.9*	339

Table 15. MT-NLG Datasets Summary. Disclosed in **bold**. Determined in *italics*.

Alan D. Thompson. 2022. *What's in my AI? A Comprehensive Analysis of Datasets used to Train GPT...*
https://LifeArchitect.ai/whats-in-my-ai

9. Gopher Datasets

The Gopher model was released by DeepMind in 2021, with 280B parameters. The paper was clear about the high-level token counts and sizes[28] of the training dataset used, but not about the detailed contents.

Count	Dataset	Raw Size (GB)	Tokens
1	**MassiveWeb**	**1900**	**506B**
2	**Books**	**2100**	**560B**
3	**C4**	**750**	**182B**
4	**News**	**2700**	**676B**
5	**GitHub**	**3100**	**422B**
6	**Wikipedia**	**1**	**4B**
	Total	10551	2350B

Table 16. Disclosed Gopher Datasets (MassiveText). Disclosed in **bold**. Determined in *italics*.

Of interest, the Gopher paper disclosed that its Books dataset contains some books that are more than 500 years old (1500-2008).

9.1. MassiveWeb Dataset Analysis

DeepMind was acquired by Google in 2014, and has access to enormous amounts of data in the creation of MassiveText. While the subset MassiveWeb is not detailed much further in the Gopher paper, an Appendix on pp 44, Figure A3b notes the Top 20 domains appearing in MassiveWeb[29]. Given the disclosed percentage represented for each domain, we can use the MassiveWeb total token count (506B tokens) and total Raw Size (1900GB) to determine the token count and size of each domain.

Count	Domain	Percentage tokens	Tokens (PT x 506B)	Size (PT x 1900GB)
1	**ScienceDirect**	**1.85%**	*9.4B*	*35.2GB*
2	**Gale**	**1.79%**	*9.1B*	*34.0GB*
3	**NCBI**	**1.59%**	*8.0B*	*30.2GB*
4	**Facebook**	**1.10%**	*5.6B*	*20.9GB*
5	**Issuu**	**0.98%**	*5.0B*	*18.6GB*
6	**Academia**	**0.93%**	*4.7B*	*17.7GB*
7	**Quora**	**0.75%**	*3.8B*	*14.3GB*
8	**Springer**	**0.73%**	*3.7B*	*13.9GB*
9	**YouTube**	**0.73%**	*3.7B*	*13.9GB*

[28] Gopher paper: https://arxiv.org/abs/2112.11446 pp 7: list of sizes and tokens.
[29] Gopher paper: https://arxiv.org/abs/2112.11446 pp 44, Figure A3b.

Count	Domain	Percentage tokens	Tokens (PT x 506B)	Size (PT x 1900GB)
10	**ProQuest Search**	**0.68%**	*3.4B*	*12.9GB*
11	**English Wikipedia**	**0.66%**	*3.3B*	*12.5GB*
12	**SlideShare**	**0.58%**	*2.9B*	*11.0GB*
13	**SlidePlayer**	**0.57%**	*2.9B*	*10.8GB*
14	**Reddit**	**0.51%**	*2.6B*	*9.7GB*
15	**Medium**	**0.42%**	*2.1B*	*8.0GB*
16	**Wiley Online Library**	**0.38%**	*1.9B*	*7.2GB*
17	**Europe PubMed Central**	**0.38%**	*1.9B*	*7.2GB*
18	**GitHub**	**0.33%**	*1.7B*	*6.3GB*
19	**DocPlayer**	**0.32%**	*1.6B*	*6.1GB*
20	**StackOverflow**	**0.28%**	*1.4B*	*5.3GB*

Table 17. MassiveWeb: Top 20 Domains. Disclosed in **bold**. Determined in *italics*.

9.2. Gopher: Concerns with Dataset Analysis of Wikipedia

The total size of the Wikipedia dataset is challenging to determine. In the Gopher paper, the researchers note that no deduplication is applied to Wikipedia[30]. However, the disparate sizes listed in the paper (12.5GB MassiveWeb Wikipedia vs 1GB MassiveText Wikipedia) may be due to an error, perhaps listing '1GB' instead of '10GB'. In any case, for this article, only the MassiveWeb dataset version (12.5GB) is used.

9.3. Gopher: No WebText

The WebText dataset of outbound Reddit links is not included as part of the Gopher dataset. For clarity, while Reddit is a Top Domain in MassiveWeb, the dataset is only capturing Reddit links within the Reddit domain. By definition, WebText[31] is comprised of "all outbound links from Reddit" (that is, links that lead to domains outside of the Reddit domain).

[30] Gopher paper: pp41n14 "Note that we apply document deduplication to all *MassiveText* subsets with the exception of Wikipedia and GitHub."
[31] GPT-2 paper, pp3.

9.4. Gopher Grouped Datasets

MassiveWeb is considered as a sub-component of MassiveText, and integrated into the dataset summary of Gopher, with grouping based on available information listed below:

Count	Source	Domain	Tokens (PT x 506B)	Size (PT x 1900GB)
1	MassiveWeb	**Wikipedia**	*3.3B*	*12.5GB*
2	MassiveText	**Books**	**560B**	**2100GB**
3	MassiveWeb	**ScienceDirect**	*9.4B*	*35.2GB*
4	MassiveWeb	**Gale**	*9.1B*	*34.0GB*
5	MassiveWeb	**NCBI**	*8.0B*	*30.2GB*
6	MassiveWeb	**Academia**	*4.7B*	*17.7GB*
7	MassiveWeb	**Springer**	*3.7B*	*13.9GB*
8	MassiveWeb	**ProQuest Search**	*3.4B*	*12.9GB*
9	MassiveWeb	**Wiley Online Library**	*1.9B*	*7.2GB*
10	MassiveWeb	**Europe PubMed Central**	*1.9B*	*7.2GB*
11	MassiveWeb	**DocPlayer**	*1.6B*	*6.1GB*
		Journals total	*41.8B*	*164.4*
12	MassiveText	**C4**	**182B**	**750GB**
13	MassiveText	**News**	**676B**	**2700GB**
		Common Crawl Total	*858B*	*3450GB*
14	MassiveText	GitHub	**422B**	**3100GB**
15	MassiveWeb	Remainder: (Size: 1900 - 12.5 - 164.4) (Tokens: 506B - 3.3 - 41.8)	*460.9B*	*1723.1GB*
		Other total	*882.9B*	*4823.1*
		Total	*2346B*	*10550GB*

Table 18. Gopher Grouped Datasets. Disclosed in **bold**. Determined in *italics*.

9.5. Gopher Dataset Summary

Gopher features the largest dataset in this article, at 10.5TB. The final dataset summary analysis of the Gopher model is:

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	**12.5**	**2100**	*164.4*		*3450*	*4823*	10550
Tokens	**3**	**560**	*42*		*858*	*883*	2346

Table 19. Gopher Datasets Summary. Disclosed in **bold**. Determined in *italics*.

10. Conclusion

We present possibly the most comprehensive synthesis and analysis of datasets used to train modern transformer large language models to early 2022. Where primary sources were opaque, the research synthesized here was collected from secondary and tertiary sources, and often necessitated assumptions to determine final estimates. As researchers approach quadrillions of tokens (1,000 trillion), and petabytes of data (1,000TB), it is becoming increasingly important to ensure that the documentation of dataset composition is disclosed in detail.

Of particular concern is the rapid progress of verbose and anonymous output from powerful AI systems based on large language models, many of which have little documentation of dataset details.

Researchers are strongly encouraged to employ templates provided in the 'Datasheet for Datasets' paper highlighted, and to use best-practice papers (i.e. The Pile v1 paper, with token count) when documenting datasets. Metrics for dataset size (GB), token count (B), source, grouping, and other details should be fully documented and published.

As language models continue to evolve and penetrate all human lives more fully, it is useful, urgent, and necessary to ensure that dataset details are accessible, transparent, and understandable for all.

11. Further reading

For brevity and readability, footnotes were used in this article, rather than in-text/parenthetical citations. Primary reference papers are listed below, or please see http://lifearchitect.ai/papers/ for the major foundational papers in the large language model space. Papers below are shown in order of appearance in this article.

Datasheets for Datasets
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for Datasets.* https://arxiv.org/abs/1803.09010

GPT-1 paper
Radford, A., & Narasimhan, K. (2018). *Improving Language Understanding by Generative Pre-Training.* OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

GPT-2 paper
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-3 paper
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., & Dhariwal, P. et al. (2020). OpenAI. *Language Models are Few-Shot Learners.* https://arxiv.org/abs/2005.14165

The Pile v1 paper
Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., & Foster, C. et al. (2021). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling.* EleutherAI. https://arxiv.org/abs/2101.00027

GPT-J announcement
Komatsuzak, A., Wang, B. (2021). *GPT-J-6B: 6B JAX-Based Transformer.* https://arankomatsuzaki.wordpress.com/2021/06/04/gpt-j/

GPT-NeoX-20B paper
Black, S., Biderman, S., Hallahan, E. et al. (2022). EleutherAI. *GPT-NeoX-20B: An Open-Source Autoregressive Language Model.* http://eaidata.bmk.sh/data/GPT_NeoX_20B.pdf

RoBERTa paper
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., & Chen, D. et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Meta AI. https://arxiv.org/abs/1907.11692

MT-NLG paper
Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., & Casper, J. et al. (2021). Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. Microsoft/NVIDIA. https://arxiv.org/abs/2201.11990

Gopher paper
Rae, J., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., & Song, F. et al. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. DeepMind. https://arxiv.org/abs/2112.11446

Appendix A: Top 50 Resources: Wikipedia + CC + WebText (i.e. GPT-3)

Based on determinations made in this article, especially token counts per resource in each dataset, we can show the rankings of top resources or domains for models which use a combination of Wikipedia + Common Crawl + WebText datasets as part of their overall training dataset. For clarity, this includes the following models: OpenAI GPT-3, EleutherAI GPT-J, EleutherAI GPT-NeoX-20B, Meta AI Megatron-11B and RoBERTA, and Microsoft/NVIDIA MT-NLG, and others.

Note that the ranking shown is based on unweighted total tokens available within datasets, and subjective weightings per dataset are calculated by researchers prior to model pre-training. Some duplication appears (e.g. The New York Times appears in both WebText at 111M tokens and filtered Common Crawl at 100M tokens).

Rank	Resource/Domain	Dataset Group	Tokens (M) Unweighted
1	Biography	Wikipedia	834
2	Google Patents	Common Crawl	750
3	Geography	Wikipedia	531
4	Google	WebText	514
5	Culture and Arts	Wikipedia	474
6	History	Wikipedia	297
7	Biology, Health, and Medicine	Wikipedia	234
8	Archive	WebText	199
9	Sports	Wikipedia	195
10	Blogspot	WebText	152
11	Business	Wikipedia	144
12	GitHub	WebText	138
13	Other society	Wikipedia	132
14	The NY Times	WebText	111
15	WordPress	WebText	107
16	Science & Math	Wikipedia	105
17	WashingtonPost	WebText	105
18	Wikia	WebText	104
19	BBC	WebText	104
20	The NY Times	Common Crawl	100
21	Los Angeles Times	Common Crawl	90
22	The Guardian	Common Crawl	90
23	PLoS	Common Crawl	90
24	TheGuardian	WebText	82
25	Forbes	Common Crawl	80
26	HuffingtonPost	Common Crawl	75
27	Patents.com	Common Crawl	71
28	Scribd	Common Crawl	70
29	eBay	WebText	70

Rank	Resource/Domain	Dataset Group	Tokens (M) Unweighted	
30	Pastebin	WebText	70	
31	CNN	WebText	66	
32	Washington Post	Common Crawl	65	
33	Yahoo	WebText	65	
34	HuffingtonPost	WebText	62	
35	Go	WebText	62	
36	The Motley Fool	Common Crawl	61	
37	Reuters	WebText	61	
38	IMDb	WebText	61	
39	IPFS	Common Crawl	60	
40	Frontiers Media	Common Crawl	60	
41	Business Insider	Common Crawl	60	
42	Chicago Tribune	Common Crawl	59	
43	Booking.com	Common Crawl	58	
44	The Atlantic	Common Crawl	57	
45	Springer Link	Common Crawl	56	
46	Al Jazeera	Common Crawl	55	
47	Kickstarter	Common Crawl	54	
48	Goo	WebText	54	
49	FindLaw Caselaw	Common Crawl	53	
50	NCBI	Common Crawl	53	
		Total	7300M	(7.3B)