A FineWeb Datasheet

Dataset Details	
Purpose of the dataset	We released FineWeb to make large language model training more accessible to the machine learning community at large.
Curated by	The dataset was curated by Hugging Face.
Funded by	The dataset was funded by Hugging Face.
Language(s)	English
License	The dataset is released under the Open Data Commons Attribution License (ODC-By) v1.0 license. The use of this dataset is also subject to CommonCrawl's Terms of Use.
Dataset Structure	
Data Instances	The following is an example sample from the dataset. It is part of the `CC-MAIN-2021-43` snapshot and was crawled on `2021-10-15T21:20:12Z`:

```
{
   "text": "This is basically a peanut flavoured cream thickened with egg yolks and then set into a ramekin on top of some jam. Tony, one of the Wedgwood chefs, suggested sprinkling on some toasted crushed peanuts at the end to create extra crunch, which I thought was a great idea. The result is excellent.",
   "id": "<urn:uuid:e5a3e79a-13d4-4147-a26e-167536fcac5d>",
   "dump": "CC-MAIN-2021-43",
   "url": "<http://allrecipes.co.uk/recipe/24758/peanut-butter-and-jam-creme-brulee.aspx?o_is=SimilarRecipes&o_ln=SimRecipes_Photo_7>",
   "date": "2021-10-15T21:20:12Z",
   "file_path": "s3://commoncrawl/crawl-data/CC-MAIN-2021-43/segments/1634323583083.92/warc/CC-MAIN-20211015192439-20211015222439-00600.warc.gz",
   "language": "en",
   "language_score": 0.948729,
   "token_count": 69
}
```

Data Fields	- `text (string)`: the main text content - `id (string)`: original unique identifier for this sample from CommonCrawl - `dump (string)`: the CommonCrawl dump/snapshot this sample was a part of - `url (string)`: url to the original page where text was present - `date (string)`: crawl date (from CommonCrawl) - `file_path (string)`: s3 path for the individual CommonCrawl warc file containing this sample - `language (string)`: en for all the samples in this dataset - `language_score (float)`: language prediction score (0.01.0) as reported by the fastText language classifier - `token_count (int)`: number of tokens when applying the gpt2 tokenizer to this sample
Data Splits	The default subset includes the entire dataset. We also include separate splits for each CommonCrawl dump. FineWeb-Edu, a subset filtered for educational content, is also available.
Dataset Creation	
Curation Rationale	With FineWeb, we aim to provide the open source community with a clean and large-scale dataset for pretraining performant large language models.
Source Data	The source data consists of webpages crawled by the CommonCrawl foundation over the 2013-2024 time period. We then extracted the main page text from the HTML of each webpage, filtered each sample and deduplicated each individual CommonCrawl dump/crawl.

Data processing steps	The data processing pipeline consists of:
	`- URL filtering`
	`- Trafilatura text extraction`
	`- FastText language filter`
	`- MassiveText repetition and quality`
	`↪ filters`
	`- C4 quality filters`
	`- FineWeb custom filters`
	`- MinHash deduplication`
	`- PII reformatting`
	For FineWeb-Edu, we further apply a filtering step based on our educational content classifier.

Annotations	We augment the original samples with the `language`, `language_score` and `tokens_count` annotations. The language related annotations are automatically generated by our language filter. `token_count` is generated by applying the GPT-2 tokenizer to the text column.
Personal and Sensitive Information	We anonymize email addresses and public IP addresses using regex patterns.

Considerations for Using the Data	
Social Impact of Dataset	With the release of FineWeb, we aim to make LLM training more accessible to the machine learning community by: (a) making the dataset creation process more transparent, by sharing our entire processing setup including the codebase used (b) helping alleviate the costs of dataset curation, both in time and in compute, for model creators by publicly releasing our dataset with the community.
Biases	Efforts were made to minimize the amount of NSFW and toxic content present in the dataset by employing filtering on the URL level. However, there are still a significant number of documents present in the final dataset that could be considered to be toxic or contain harmful content. As FineWeb was sourced from the web as a whole, any harmful biases typically present in the web may be reproduced on our dataset. Bias analyses for sensitive subgroups demonstrate that 'man' is more common in the dataset than other gender terms, 'christian' is more common than other religion terms. The disproportionate association of specific terms to sensitive subgroups is relatively low, with the most notable bias that some religion terms tend to be more associated with online dating terms. We provide a more detailed bias analysis in Section 5.
Other Known Limitations	As a consequence of some of the filtering steps applied, it is likely that code content is not prevalent in our dataset. Users are advised to consider complementing FineWeb with other code datasets and specialized curated sources, such as Wikipedia, which may have better formatting than the Wikipedia content included in FineWeb.

B License and hosting

The FineWeb datasets are released under the Open Data Commons Attribution License (ODC-By) v1.0. The full text of the license is available at `https://opendatacommons.org/licenses/by/1-0/`. The use of the dataset is also subject to CommonCrawl's Terms of Use. The authors of this work are solely responsible for the content and the views presented herein. NeurIPS is not associated and shall bear no responsibility for the work presented, including the dataset itself.

The FineWeb datasets are hosted on the HuggingFace hub, where they will remain available for the foreseeable future. We plan to regularly update the dataset with new CommonCrawl snapshots as they are released.

C Linked resources

Resource	URL
FineWeb repository (DOI `10.57967/hf/2493`)	https://hf.co/datasets/HuggingFaceFW/fineweb
FineWeb Croissant metadata	https://hf.co/api/datasets/HuggingFaceFW/fineweb/croissant
FineWeb-Edu repository (DOI `10.57967/hf/2497`)	https://hf.co/datasets/HuggingFaceFW/fineweb-edu
FineWeb-Edu Croissant metadata	https://hf.co/api/datasets/HuggingFaceFW/fineweb-edu/croissant
FineWeb Llama3 annotations	https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu-llama3-annotations
Educational classifier	https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier
Dataset comparison models	https://hf.co/collections/HuggingFaceFW/comparison-models-662457b0d213e8c14fe47f32
Ablation models	https://hf.co/collections/HuggingFaceFW/data-experiments-665ed849020d8b66a5d9896f
Datatrove processing code to reproduce FineWeb	https://github.com/huggingface/datatrove/blob/main/examples/fineweb.py
Evaluation setup	https://hf.co/datasets/HuggingFaceFW/fineweb/blob/main/lighteval_tasks.py

D Data ablation setup

D.1 Model architecture

Parameter	Value
Architecture	Llama
Number of attention heads	32
Number of hidden layers	24
Number of key-value heads	32
RMS Norm epsilon	1e-05
Tied word embeddings	True
Embedding size	50257
Total number of parameters	1.71B
Random initialization std	0.02
Tokenizer	GPT2

D.2 Distributed training setup

Parameter	Value
Data parallelism (dp)	64
Tensor parallelism (tp)	1
Pipeline parallelism (pp)	1
Micro-batch size	4
Sequence length	2048
Batch accumulation per replica	4

D.3 Optimizer Configuration

Parameter	Value
Adam beta1	0.9
Adam beta2	0.95
Adam epsilon	1.0e-8
Gradient clipping	1.0
Weight decay	0.1
Learning rate	3e-4
Warmup steps	500
Warmup style	linear
Decay style	cosine
Minimum decay LR	3.0e-5

E Deduplication

E.1 Deduplication parameters

As mentioned in Section 3.4, we use 5-grams and 112 hash functions for our MinHash deduplication. Each 5-gram is hashed with each of the 112 hash functions, and a document signature is obtained by taking the minimum hash value (minhash) across all 5-grams for each hash function. We further split the resulting 112 minhashes into 14 buckets of 8 hashes each. Documents are matched if they have the same 8 minhashes in at least one of the 14 buckets.

With these parameters, the probability that two documents with a n-gram similarity (s) of 0.7, 0.75, 0.8 and 0.85 would be identified as duplicates would be 56%, 77%, 92% and 98.8%, respectively. This split therefore will match documents that are at least 75% similar with a high probability, and almost guarantee that documents with similarities of 85% or above will be matched. These values can be computed by taking the following probabilities: that the two documents would have the same value for a given hash function, s; that they do not have the same 8 minhashes in one bucket, $1 - s^8$; that they do not have the same 8 minhashes in any of the 14 buckets, $(1 - s^8)^{14}$; and finally that they have the same 8 minhashes on at least one of the 14 buckets, $1 - (1 - s^8)^{14}$.

See Fig. 13 for a match probability comparison between our setup with 112 hashes and the one from RefinedWeb, with 9000 hashes, divided into 450 buckets of 20 hashes.

While the high number of hash functions in RefinedWeb allows for a steeper, more well-defined cut off (document pairs with similarity near the threshold are more likely to be correctly identified), this

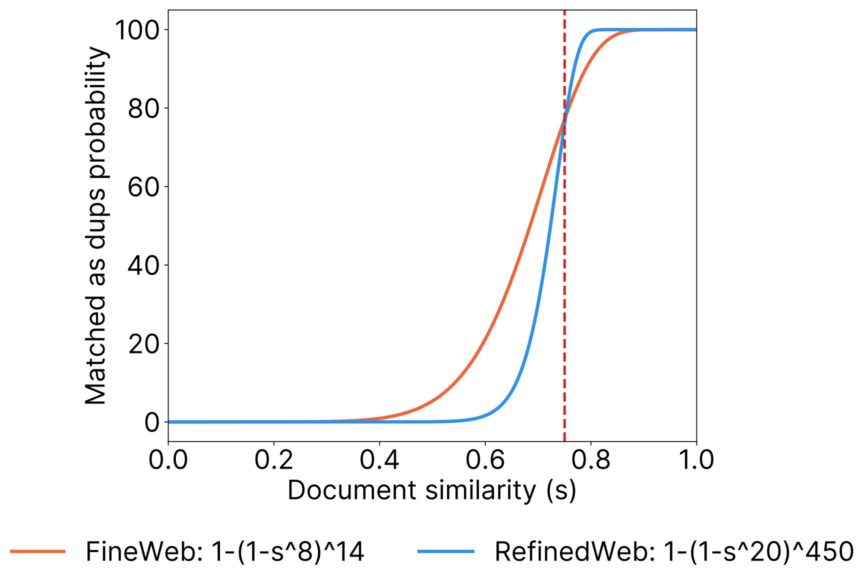

Figure 13: Comparison between FineWeb and RefinedWeb document matching probabilities.

larger number of hash functions also requires a substantially larger amount of compute resources, as each individual hash must be computed, stored, and then compared with hashes from other documents. We believe the compute and storage savings make up for the higher uncertainty on documents near the threshold.

E.2 Measuring the effect of deduplication

Given the nature of deduplication, its effect is not always visible in a smaller slice of the dataset (such as 28B tokens, the size used for our filtering ablations). Furthermore, there are specific effects at play when deduplicating across different Common Crawl dumps, as some URLs and webpages are recrawled from one snapshot to the next.

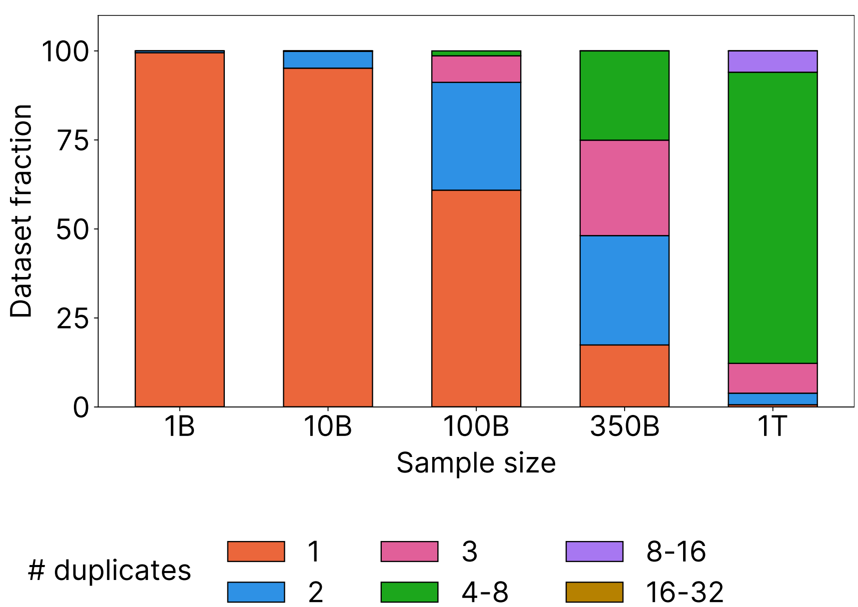

Figure 14: **Small ablations are ineffective for deduplication analysis.** The chart displays the distribution of document repetitions across different sample sizes (1 billion, 10 billion, 100 billion, 350 billion, and 1 trillion tokens) from a dataset of 20T tokens.

To visualize the effect of scaling the number of training tokens when measuring deduplication impact, we simulated creating different-sized subsets of randomly sampled documents from the full dataset under the following extreme conditions: there are 100 snapshots, where each one is made up of unique documents with a total of 200 billion tokens (yielding our total of 20 trillion from Section 3.4), and each snapshot is an exact copy of each other (worst case scenario for inter snapshot duplication).

In Fig. 14, we can see that for a 1 billion subset, almost all documents would be unique (#duplicates=1), despite each document being repeated 100 times in the full dataset. At the 100 billion scale (0.5% of the total dataset), there starts to be a larger number of documents being repeated twice, and a few even 4-8 times. At the larger scale of 1 trillion (5% of the total dataset), the majority of the documents are repeated up to 8 times, with some being repeated up to 16 times. This simulation illustrates the inherent difficulties with measuring deduplication impact on the training of larger LLMs once the largest duplicate clusters have been removed. We ran our performance evaluations for deduplicated data at the 350 billion scale, which would, under this theoretical scenario, be made up of a significant portion of documents duplicated up to 8 times.

E.3 Alternative global deduplication

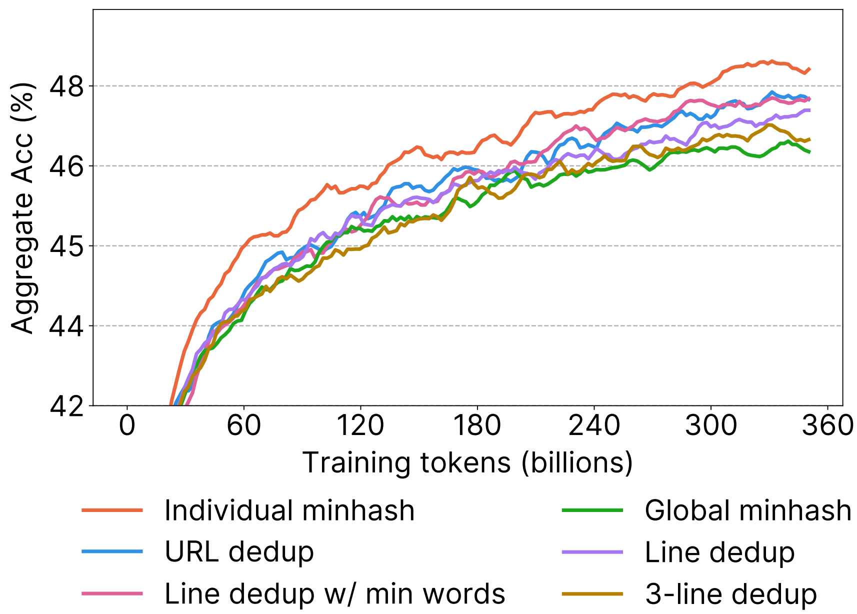

Figure 15: **URL and Line-wise deduplication study**. None of the attempted deduplication methods outperform individual deduplication.

To attempt to improve performance on top of independently deduplicating each snapshot, we experimented with applying other "lighter" global deduplication methods to all the individually MinHash deduplicated snapshots (comprising 20 trillion tokens of data).

We explored URL deduplication, where we only kept one document per normalized (lowercased) URL (71.5% of tokens removed, 5.6 trillion left) — *FineWeb URL dedup*. Different line-based deduplication variations were also considered: remove all but 1 (randomly chosen) occurrence of each duplicated line (77.8% of tokens dropped, 4.4 trillion left) — *FineWeb line dedup*; same as above, but only removing duplicate lines with at least 10 words and dropping documents with fewer than 3 sentences after deduplication (85% of tokens dropped, 2.9 trillion left) — *FineWeb line dedup w/ min words*; and remove all but 1 occurrence of each span of 3 duplicated lines with each number treated as 0 when finding duplicates, (80.9% of tokens removed, 3.7 trillion left) — *FineWeb 3-line dedup*.

As can be seen in Fig. 15 the performance of the models trained on each of these methods was consistently worse (albeit to different degrees) than that of the original individually deduplicated data. We therefore did not apply any additional deduplication beyond individual-snapshot MinHash-based deduplication.

E.4 Other filters considered

Metric	Threshold	Aggregate Acc (%)	Tokens removed (%)
lines-with-punct-ratio	≥ 0.12	42.85	10.14
duplicated-line-char-ratio	≤ 0.01	42.78	12.47
lines-with-punct-ratio	≥ 0.12 or $= 0$	42.72	5.82
lines-shorter-30-ratio	≤ 0.67	42.65	3.37
line-with-most-3-words-ratio	≤ 0.49	42.61	2.51
duplicate-(5-10)-grams-char-ratio	≤ 0.1, 0.084, 0.073, 0.065, 0.057, 0.05	42.60	10.92
lines-with-punct-ratio	≥ 0.08 or $= 0$	42.59	3.42
top-(2,3,4)-gram-char-ratio	$\leq$ 0.13, 0.087, 0.079	42.58	56.71
lines-shorter-30-ratio	0.69	42.58	3.73
avg-words-per-line	≥ 7	42.56	2.32
lines-shorter-30-ratio	≤ 0.5	42.53	11.17
avg-words-per-line	≥ 5	42.39	0.83
avg-words-per-line	≥ 9	42.27	4.47
avg-line-length-0.5-sampling	≥ 56	42.93	3.24
avg-line-length	≥ 56	42.12	6.48
avg-line-length-0.5-sampling	≥ 40	42.03	1.50

Table 2: Full list of heuristic filters tested

F FineWeb-Edu

F.1 Annotation Prompt

We use the following prompt template to generate document annotations using the Llama3 model:

> Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:
>
> - Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
> - Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
> - Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
> - Grant a fourth point if the extract highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
> - Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.
>
> The extract: <EXAMPLE>.
>
> After examining the extract:
>
> - Briefly justify your total score, up to 100 words.
> - Conclude with the score using the format: "Educational score: <total points>"

F.2 Additional results

Fig. 16 compares FineWeb-Edu to other open web datasets on 9 becnhmarks, using a 1.71B model trained on 350 billion tokens. Additionally, Fig. 17 displays the results of experiments with various filtering thresholds for building FineWeb-Edu, using a 1.71B model trained on 28 billion tokens. Our findings indicate that a threshold of 3 yields the best average performance.

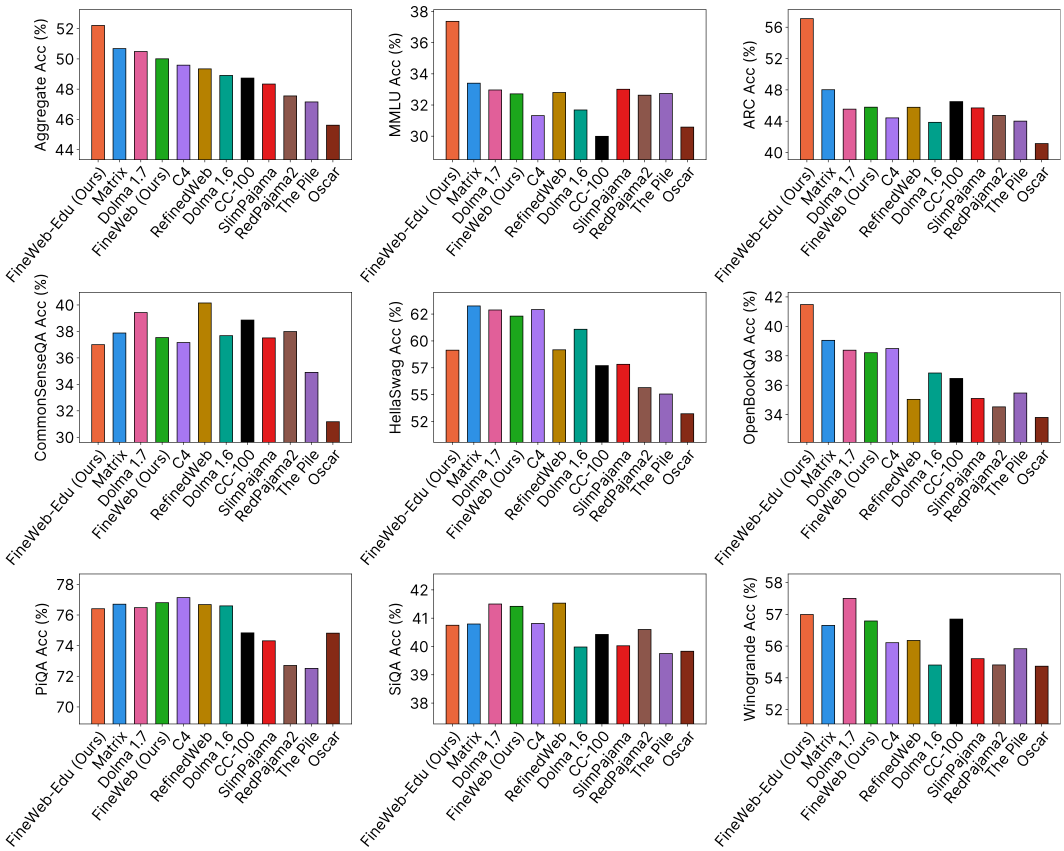

Figure 16: **Comparing FineWeb datasets to other public datasets on each benchmark.**

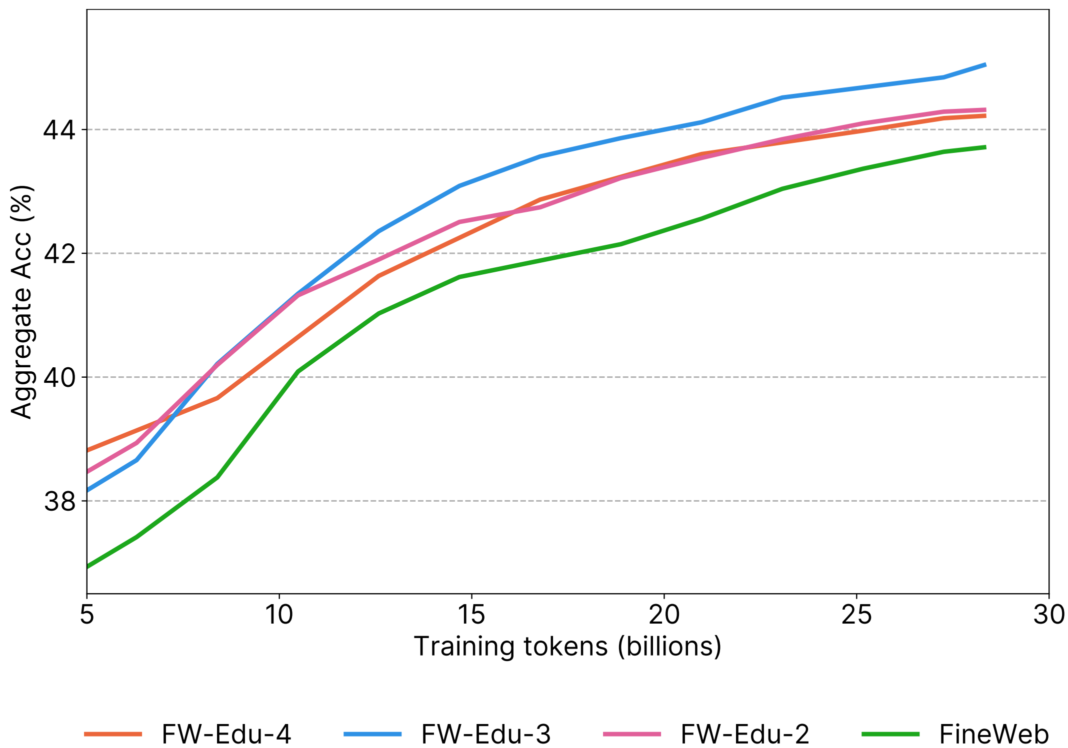

Figure 17: **Ablation study of FineWeb Edu thresholds**. Using a filtering threshold of 3 yields the best Aggregate Accuracy when building FineWeb-Edu. FW-Edu-i denotes dataset filtered to only contain documents with an educational score greater or equal i.

F.3 Topic distribution

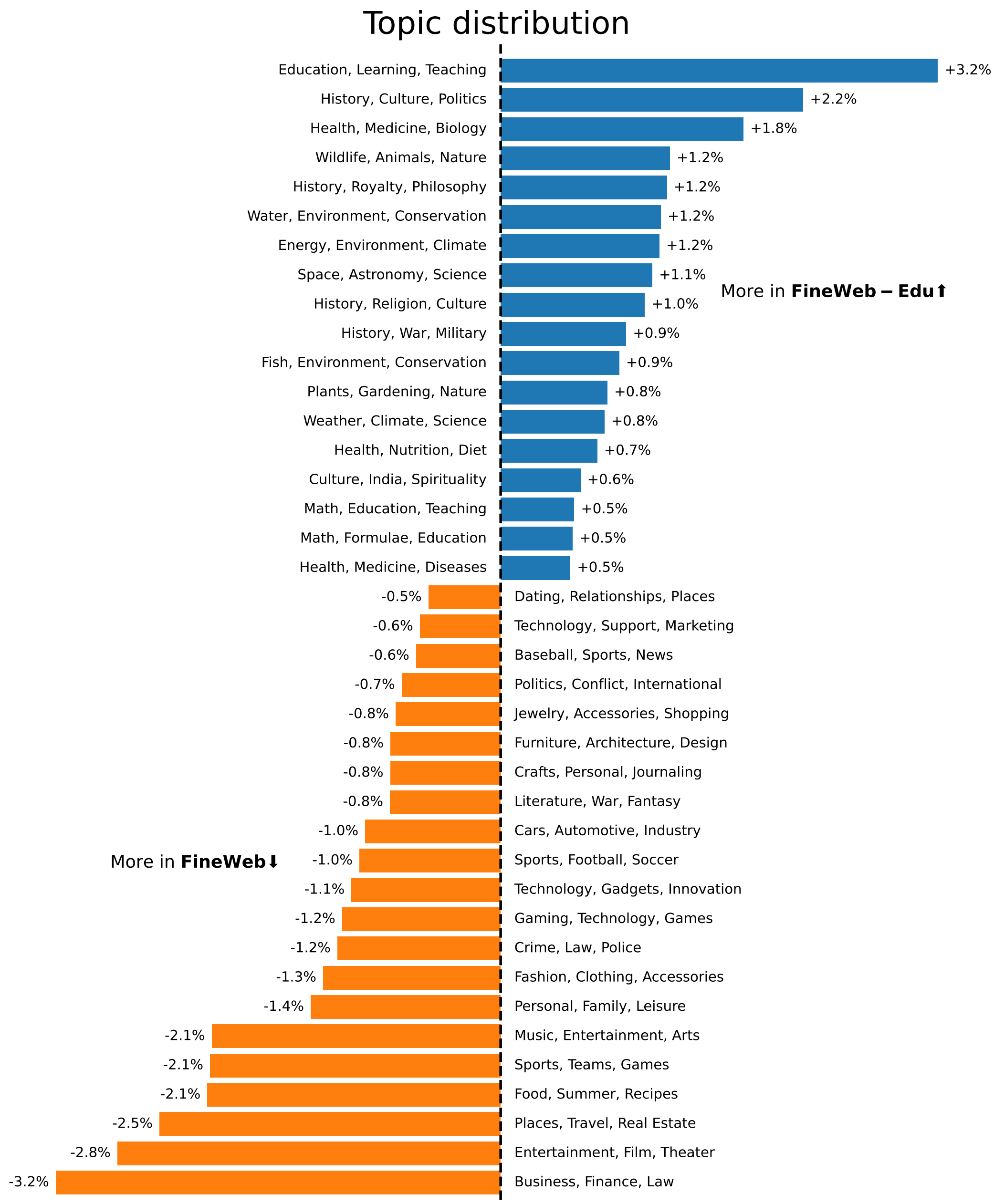

Figure 18: **FineWeb and FineWeb-Edu topic comparison**. FineWeb-Edu has a higher representation of topics like 'Education, Learning, Teaching' and 'History, Culture, Politics' compared to FineWeb. Conversely, it down-samples topics such as 'Business, Finance, Law' and 'Entertainment, Film, Theater.' Values indicate the absolute difference in the percentage of each topic between the datasets and only topics with an absolute difference of at least 0.5% are displayed.

F.4 Domain fit

Source	Domain	FineWeb ppl	FineWeb-Edu ppl
Dolma V1.5	common-crawl	**14.499**	18.336
Dolma V1.5	pes2o	12.226	**10.242**
Dolma V1.5	reddit uniform	**23.814**	29.864
Dolma V1.5	stack uniform	7.65	**7.014**
Dolma V1.5	wiki	**12.0**	12.243
M2D2 Wikipedia	Culture and the arts	**10.367**	14.518
M2D2 Wikipedia	Culture and the arts Culture and Humanities	**14.037**	14.116
M2D2 Wikipedia	Culture and the arts Games and Toys	**15.774**	18.912
M2D2 Wikipedia	Culture and the arts Mass media	**14.352**	18.134
M2D2 Wikipedia	Culture and the arts Performing arts	14.311	**13.313**
M2D2 Wikipedia	Culture and the arts Sports and Recreation	**11.295**	14.735
M2D2 Wikipedia	Culture and the arts The arts and Entertainment	**13.669**	19.039
M2D2 Wikipedia	Culture and the arts Visual arts	**14.967**	15.158
M2D2 Wikipedia	General referece	11.962	**11.246**
M2D2 Wikipedia	General referece Further research tools and topics	**16.202**	19.191
M2D2 Wikipedia	General referece Reference works	**14.914**	18.621
M2D2 Wikipedia	Health and fitness	**12.0**	13.448
M2D2 Wikipedia	Health and fitness Exercise	**11.874**	13.951
M2D2 Wikipedia	Health and fitness Health science	11.509	**10.997**
M2D2 Wikipedia	Health and fitness Human medicine	**12.0**	13.448
M2D2 Wikipedia	Health and fitness Nutrition	10.09	**8.489**
M2D2 Wikipedia	Health and fitness Public health	12.804	**11.797**
M2D2 Wikipedia	Health and fitness Self care	14.62	**12.782**
M2D2 Wikipedia	History and events	13.446	**12.516**
M2D2 Wikipedia	History and events By continent	14.174	**12.066**
M2D2 Wikipedia	History and events By period	12.94	**11.0**
M2D2 Wikipedia	History and events By region	13.61	**11.63**
M2D2 Wikipedia	Human activites	**15.159**	18.728
M2D2 Wikipedia	Human activites Human activities	12.784	**11.117**
M2D2 Wikipedia	Human activites Impact of human activity	15.092	**13.592**
M2D2 Wikipedia	Mathematics and logic	12.703	**9.903**
M2D2 Wikipedia	Mathematics and logic Fields of mathematics	12.703	**9.903**
M2D2 Wikipedia	Mathematics and logic Logic	14.281	**13.367**
M2D2 Wikipedia	Mathematics and logic Mathematics	14.923	**14.207**
M2D2 Wikipedia	Natural and physical sciences	12.884	**10.529**
M2D2 Wikipedia	Natural and physical sciences Biology	12.718	**10.221**
M2D2 Wikipedia	Natural and physical sciences Earth sciences	15.346	**13.145**

Source	Domain	FineWeb ppl	FineWeb-Edu ppl
M2D2 Wikipedia	Natural and physical sciences Nature	12.594	**9.886**
M2D2 Wikipedia	Natural and physical sciences Physical sciences	13.088	**10.643**
M2D2 Wikipedia	Philosophy and thinking	**14.081**	16.067
M2D2 Wikipedia	Philosophy and thinking Philosophy	14.209	**12.91**
M2D2 Wikipedia	Philosophy and thinking Thinking	**14.081**	16.067
M2D2 Wikipedia	Religion and belief systems	12.636	**11.326**
M2D2 Wikipedia	Religion and belief systems Allah	14.072	**10.808**
M2D2 Wikipedia	Religion and belief systems Belief systems	12.843	**11.652**
M2D2 Wikipedia	Religion and belief systems Major beliefs of the world	13.824	**11.834**
M2D2 Wikipedia	Society and social sciences	11.777	**11.195**
M2D2 Wikipedia	Society and social sciences Social sciences	**11.81**	13.03
M2D2 Wikipedia	Society and social sciences Society	11.777	**11.195**
M2D2 Wikipedia	Technology and applied sciences	11.592	**9.368**
M2D2 Wikipedia	Technology and applied sciences Agriculture	**13.941**	14.998
M2D2 Wikipedia	Technology and applied sciences Computing	**15.562**	16.091
M2D2 Wikipedia	Technology and applied sciences Engineering	14.897	**13.861**
M2D2 Wikipedia	Technology and applied sciences Transport	**16.519**	17.886
Manosphere	avfm	**27.332**	32.058
Manosphere	incels	**18.253**	20.788
Manosphere	love shy	**28.206**	33.374
Manosphere	mgtow	**24.913**	29.702
Manosphere	pua forum	**25.133**	33.297
Manosphere	red pill talk	**33.87**	42.947
Manosphere	reddit	**24.786**	30.903
Manosphere	rooshv	**23.593**	27.819
Manosphere	the attraction	**24.988**	30.907
RedPajama	arxiv	32.338	**23.368**
RedPajama	books	**22.095**	23.953
RedPajama	c4	**12.685**	15.599
RedPajama	commoncrawl	**8.0**	8.979
RedPajama	github	5.613	**5.247**
RedPajama	stackexchange	9.055	**8.862**
RedPajama	wikipedia	8.741	**8.608**
Twitter AAE	AA	**246.907**	575.106
Twitter AAE	white	**98.536**	192.374

Table 3: **Paloma domain comparison between FineWeb and FineWeb-Edu**. Lower perplexity (ppl) in bold. A lower perplexity value indicates a better fit to a given domain.

G Bias Analyses

G.1 Distributional Analysis

Subgroup	Terms
age	'old', 'young'
gender	'man', 'woman', 'non-binary'
religion	'muslim', 'christian', 'jewish', 'hindu', 'buddhist', 'atheist'

Table 4: Subgroups and terms used for bias analyses.

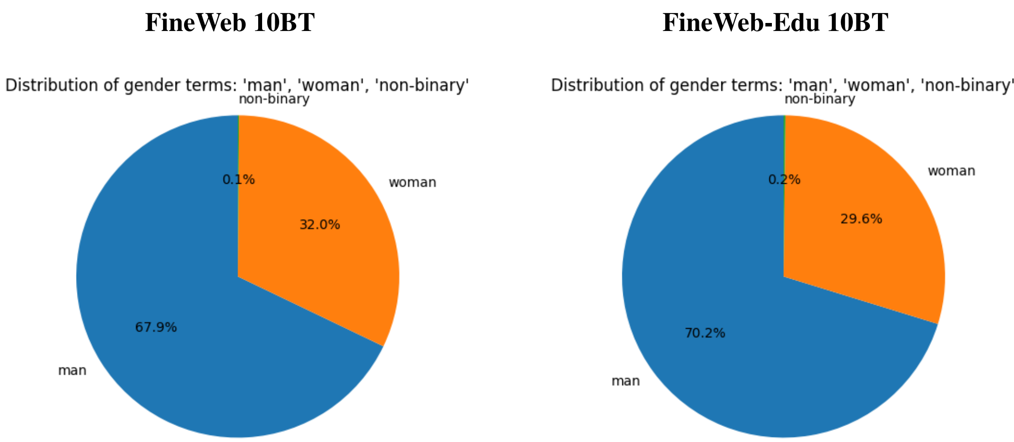

Figure 19: Distribution of *gender* terms in FineWeb (Left) and FineWeb-Edu (Right), 10BT samples.

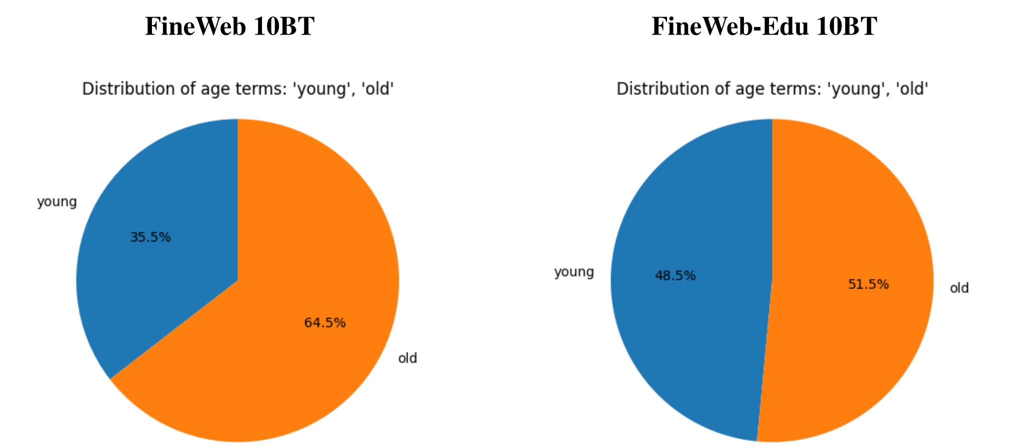

Figure 20: Distribution of *age* terms in FineWeb (Left) and FineWeb-Edu (Right), 10BT samples.

To begin, we examine the distribution over subgroup terms for *gender* (Fig. 19) *age* (Fig. 20), and *religion* (Fig. 21) in a subset of FineWeb and FineWeb-Edu randomly sampled from the whole dataset, of around 10 Billion GPT-2 tokens (FineWeb 10BT and FineWeb-Edu 10BT). Terms used are shown in Table 4 and are all normalized to lowercase for this analysis.

We find that 'man' appears much more frequently than 'woman' and 'non-binary', and 'christian' appears much more frequently than all other religions terms tested.

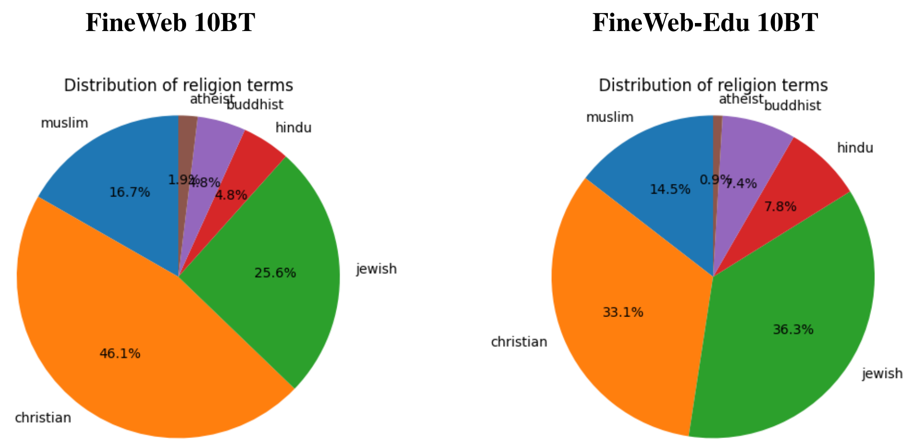

Figure 21: Distribution of *religion* terms in FineWeb (Left) and FineWeb-Edu (Right), 10BT samples.

G.2 Association Analysis

We next examine the skews with respect to the different subgroup terms, as measured by TF-IDF [78]. This method is described as capturing the *specificity* of words in the dataset, here applied as specificity with respect to the terms for the different subgroups. This provides a way to quantify how "biased" each subgroup term is with respect to the words they co-occur with. Specifically, given the dataset and terms for a subgroup of interest, we:

1. Build a vocabulary of all words that occur at least twice in the dataset.
2. Extract all data instances where the subgroup term is present.
3. Compute the TF-IDF for all words in the vocabulary that co-occur in the same documents as a given subgroup term.
4. Compute the difference between the TF-IDF for the given subgroup terms and the average TF-IDF of all other words they co-occur with.
5. Extract the words co-occurring with the given subgroup terms with a TF-IDF greater than 0.

G.2.1 Gender

We find that 'man' is associated with terms such as 'god', 'police', 'said' and 'good', 'woman' is associated with terms like 'said', 'women', 'police', 'life', 'love', 'dating' and 'family', and 'non-binary' is associated with 'gender' and LGBTQIA+ terms such as 'trans', 'transgender', and 'queer' (Fig. 22). Applying this same analysis to FineWeb-Edu-Sample-10BT, we find that 'man' is associated with the term 'god', and slightly associated with terms like 'war', 'great', and 'king'. 'woman' is associated with terms like 'pregnancy', 'cancer', 'mother', 'children', and 'family'.

G.2.2 Religion

Throughout, we see skews towards words associated with online intimacy: 'online', 'singles', 'sex', 'mature', 'girls'. As can be seen in Fig. 27, 'jewish' is particularly associated with 'dating' and 'singles'. 'muslim', 'jewish', 'hindu' and 'buddhist' are slightly skewed to co-occur with 'women', while 'sex' is skewed with 'muslim', 'christian', 'jewish'; and 'girl' with 'muslim', 'jewish', 'hindu'.

G.2.3 Age

The word 'young' is skewed to co-occur with 'women', consistent with the problematic tendencies in English-speaking societies to infantilize women and over-indexing on womens' youth [79, 80]. We also see expected skews, such as 'young' co-occurring with words like 'children' and 'school'.

word	non-binary	non-binary+	man	man+	woman	woman+
non-binary	0.092	0.061	0.000	-0.031	0.000	-0.031
gender	0.068	0.044	0.001	-0.023	0.003	-0.021
trans	0.055	0.037	0.000	-0.018	0.001	-0.018
transgender	0.035	0.023	0.000	-0.012	0.001	-0.011
queer	0.033	0.021	0.000	-0.011	0.001	-0.011
people	0.044	0.016	0.018	-0.009	0.020	-0.007
women	0.044	0.015	0.011	-0.018	0.031	0.003
lgbtq	0.020	0.013	0.000	-0.007	0.000	-0.006
community	0.020	0.011	0.004	-0.006	0.004	-0.005
sexual	0.017	0.008	0.003	-0.005	0.006	-0.003
female	0.016	0.007	0.003	-0.006	0.006	-0.002
sex	0.019	0.006	0.007	-0.006	0.012	-0.001
work	0.016	0.004	0.009	-0.003	0.010	-0.002
person	0.014	0.004	0.007	-0.003	0.008	-0.001
feel	0.012	0.003	0.006	-0.003	0.008	-0.001
dating	0.015	0.002	0.009	-0.004	0.015	0.002
ve	0.012	0.002	0.009	-0.001	0.010	-0.000
new	0.015	0.001	0.013	-0.001	0.013	-0.000
like	0.022	0.001	0.019	-0.002	0.021	0.000
men	0.015	0.001	0.013	-0.001	0.014	-0.000
want	0.012	0.001	0.009	-0.002	0.011	0.000
world	0.013	0.001	0.011	-0.000	0.011	-0.001
really	0.012	0.001	0.010	-0.001	0.011	0.000
year	0.010	0.001	0.008	-0.000	0.008	-0.001
young	0.010	0.001	0.008	-0.001	0.009	0.000

word	woman	woman+	non-binary	non-binary+	man	man+
woman	0.051	0.026	0.011	-0.013	0.011	-0.013
said	0.022	0.004	0.011	-0.007	0.022	0.004
women	0.031	0.003	0.044	0.015	0.011	-0.018
police	0.012	0.003	0.003	-0.007	0.014	0.004
life	0.017	0.002	0.011	-0.003	0.015	0.001
love	0.015	0.002	0.012	-0.001	0.012	-0.001
dating	0.015	0.002	0.015	0.002	0.009	-0.004
family	0.010	0.002	0.007	-0.002	0.009	-0.000
did	0.011	0.002	0.006	-0.003	0.012	0.002
just	0.019	0.002	0.015	-0.002	0.018	0.000
know	0.015	0.002	0.012	-0.002	0.014	0.000
time	0.017	0.001	0.013	-0.002	0.017	0.001
day	0.012	0.001	0.008	-0.002	0.011	0.001
good	0.011	0.001	0.007	-0.003	0.012	0.002
story	0.011	0.001	0.009	-0.001	0.009	-0.000
going	0.010	0.001	0.007	-0.002	0.010	0.001
say	0.010	0.001	0.008	-0.002	0.010	0.001
god	0.012	0.001	0.003	-0.008	0.018	0.007
years	0.012	0.001	0.010	-0.001	0.011	0.001
don	0.014	0.001	0.013	0.000	0.012	-0.001
book	0.010	0.001	0.009	-0.000	0.008	-0.001
right	0.009	0.001	0.008	-0.001	0.009	0.000

A **B**

word	man	man+	woman	woman+	non-binary	non-binary+
man	0.046	0.022	0.019	-0.005	0.007	-0.017
god	0.018	0.007	0.012	0.001	0.003	-0.008
police	0.014	0.004	0.012	0.003	0.003	-0.007
said	0.022	0.004	0.022	0.004	0.011	-0.007
good	0.012	0.002	0.011	0.001	0.007	-0.003
did	0.012	0.002	0.011	0.002	0.006	-0.003
say	0.010	0.001	0.010	0.001	0.008	-0.002
time	0.017	0.001	0.017	0.001	0.013	-0.002
day	0.011	0.001	0.012	0.001	0.008	-0.002
going	0.010	0.001	0.010	0.001	0.007	-0.002
years	0.011	0.001	0.012	0.001	0.010	-0.001
life	0.015	0.001	0.017	0.002	0.011	-0.003

C

Figure 22: Most skewed associations in FineWeb for *gender* terms 'non-binary' (A), 'woman' (B), and 'man' (C) in FineWeb compared to one another, measured using TF-IDF. Columns are sorted by the 'non-binary+', 'woman+' and 'man+' columns, measuring the difference from the mean over all words occurring more than once in the dataset.

word	atheist	atheist+	muslim	muslim+	christian	christian+	jewish	jewish+	hindu	hindu+	buddhist	buddhist+
atheist	0.095	0.078	0.001	-0.016	0.003	-0.014	0.001	-0.016	0.001	-0.016	0.001	-0.016
god	0.080	0.057	0.010	-0.013	0.026	0.003	0.008	-0.015	0.008	-0.014	0.006	-0.017
religion	0.043	0.030	0.010	-0.004	0.007	-0.007	0.005	-0.009	0.010	-0.004	0.008	-0.006
religious	0.039	0.028	0.007	-0.004	0.006	-0.006	0.005	-0.006	0.006	-0.005	0.006	-0.006
church	0.027	0.017	0.004	-0.006	0.016	0.006	0.005	-0.005	0.003	-0.007	0.005	-0.005
people	0.035	0.014	0.021	-0.001	0.019	-0.002	0.019	-0.002	0.016	-0.006	0.019	-0.003
think	0.021	0.012	0.008	-0.001	0.008	-0.001	0.007	-0.002	0.004	-0.005	0.007	-0.002
don	0.021	0.011	0.009	-0.001	0.009	-0.001	0.007	-0.002	0.004	-0.005	0.007	-0.002
life	0.023	0.010	0.009	-0.004	0.013	-0.000	0.010	-0.004	0.011	-0.003	0.015	0.002
like	0.024	0.009	0.014	-0.001	0.015	-0.001	0.014	-0.002	0.011	-0.004	0.014	-0.002
know	0.020	0.009	0.010	-0.001	0.011	0.000	0.009	-0.001	0.006	-0.004	0.008	-0.003
just	0.022	0.009	0.013	-0.001	0.014	-0.000	0.012	-0.002	0.009	-0.005	0.013	-0.001
world	0.020	0.008	0.012	-0.000	0.010	-0.002	0.011	-0.001	0.010	-0.003	0.010	-0.002
way	0.016	0.006	0.008	-0.001	0.009	-0.000	0.008	-0.002	0.006	-0.003	0.010	0.000
good	0.015	0.006	0.009	-0.001	0.010	0.001	0.008	-0.002	0.006	-0.003	0.008	-0.001
time	0.017	0.005	0.011	-0.001	0.011	-0.000	0.010	-0.001	0.008	-0.003	0.011	-0.000
man	0.015	0.004	0.012	0.001	0.012	0.001	0.011	0.000	0.009	-0.002	0.007	-0.004
christian	0.031	0.003	0.016	-0.012	0.077	0.049	0.018	-0.009	0.013	-0.015	0.011	-0.016
catholic	0.012	0.002	0.006	-0.004	0.015	0.005	0.014	0.003	0.006	-0.004	0.010	-0.001

Figure 23: Most skewed associations in FineWeb for 'atheist' compared to other religions, measured using TF-IDF. Columns are sorted by the 'atheist+' column, measuring the difference from the mean over all words.

word	buddhist	buddhist+	atheist	atheist+	muslim	muslim+	christian	christian+	jewish	jewish+	hindu	hindu+
buddhist	0.169	0.134	0.003	-0.033	0.006	-0.029	0.005	-0.030	0.007	-0.029	0.022	-0.013
single	0.055	0.018	0.003	-0.034	0.034	-0.003	0.033	-0.004	0.045	0.007	0.054	0.017
singles	0.085	0.015	0.002	-0.068	0.065	-0.005	0.076	0.006	0.110	0.041	0.079	0.010
personals	0.034	0.012	0.001	-0.021	0.018	-0.004	0.018	-0.004	0.032	0.009	0.031	0.009
site	0.038	0.007	0.004	-0.027	0.033	0.002	0.035	0.004	0.043	0.012	0.035	0.003
men	0.036	0.007	0.009	-0.021	0.030	0.001	0.028	-0.002	0.034	0.004	0.040	0.011
women	0.046	0.006	0.010	-0.030	0.048	0.007	0.037	-0.004	0.050	0.009	0.053	0.012
chat	0.019	0.005	0.001	-0.013	0.014	0.000	0.016	0.002	0.016	0.002	0.017	0.003
meet	0.028	0.004	0.003	-0.021	0.028	0.003	0.026	0.001	0.034	0.010	0.027	0.003
100	0.013	0.003	0.002	-0.007	0.008	-0.001	0.010	0.000	0.011	0.002	0.011	0.002
essay	0.013	0.003	0.003	-0.007	0.007	-0.003	0.009	-0.001	0.010	0.001	0.017	0.007
free	0.035	0.002	0.008	-0.024	0.034	0.002	0.038	0.005	0.042	0.009	0.038	0.005
date	0.011	0.002	0.002	-0.008	0.010	0.000	0.012	0.002	0.012	0.002	0.012	0.002
life	0.015	0.002	0.023	0.010	0.009	-0.004	0.013	-0.000	0.010	-0.004	0.011	-0.003
asian	0.011	0.001	0.001	-0.009	0.011	0.002	0.009	-0.001	0.012	0.002	0.014	0.004
looking	0.014	0.001	0.004	-0.009	0.015	0.002	0.015	0.001	0.018	0.005	0.015	0.001

Figure 24: Most skewed associations in FineWeb for 'buddhist' compared to other religions, measured using TF-IDF. Columns are sorted by the 'buddhist+' column, measuring the difference from the mean over all words.

word	christian	christian+	jewish	jewish+	hindu	hindu+	buddhist	buddhist+	atheist	atheist+	muslim	muslim+
dating	0.192	0.049	0.212	0.069	0.146	0.004	0.133	-0.009	0.009	-0.134	0.164	0.021
christian	0.077	0.049	0.018	-0.009	0.013	-0.015	0.011	-0.016	0.031	0.003	0.016	-0.012
online	0.047	0.010	0.057	0.020	0.038	0.001	0.032	-0.005	0.004	-0.033	0.043	0.006
sites	0.023	0.009	0.022	0.008	0.013	-0.002	0.009	-0.005	0.002	-0.013	0.019	0.004
singles	0.076	0.006	0.110	0.041	0.079	0.010	0.085	0.015	0.002	-0.068	0.065	-0.005
church	0.016	0.006	0.005	-0.005	0.003	-0.007	0.005	-0.005	0.027	0.017	0.004	-0.006
free	0.038	0.005	0.042	0.009	0.038	0.005	0.035	0.002	0.008	-0.024	0.034	0.002
catholic	0.015	0.005	0.014	0.003	0.006	-0.004	0.010	-0.001	0.012	0.002	0.006	-0.004
site	0.035	0.004	0.043	0.012	0.035	0.003	0.038	0.007	0.004	-0.027	0.033	0.002
love	0.017	0.003	0.014	-0.000	0.016	0.001	0.014	-0.001	0.013	-0.002	0.013	-0.001
god	0.026	0.003	0.008	-0.015	0.008	-0.014	0.006	-0.017	0.080	0.057	0.010	-0.013
chat	0.016	0.002	0.016	0.002	0.017	0.003	0.019	0.005	0.001	-0.013	0.014	0.000
best	0.015	0.002	0.016	0.003	0.014	0.001	0.013	-0.000	0.006	-0.006	0.014	0.001
date	0.012	0.002	0.012	0.002	0.012	0.002	0.011	0.002	0.002	-0.008	0.010	0.000
meet	0.026	0.001	0.034	0.010	0.027	0.003	0.028	0.004	0.003	-0.021	0.028	0.003
sex	0.015	0.001	0.015	0.002	0.014	0.000	0.011	-0.002	0.005	-0.008	0.020	0.007
looking	0.015	0.001	0.018	0.005	0.015	0.001	0.014	0.001	0.004	-0.009	0.015	0.002
gay	0.013	0.001	0.016	0.004	0.013	0.001	0.010	-0.002	0.006	-0.006	0.013	0.001
man	0.012	0.001	0.011	0.000	0.009	-0.002	0.007	-0.004	0.015	0.004	0.012	0.001
good	0.010	0.001	0.008	-0.002	0.006	-0.003	0.008	-0.001	0.015	0.006	0.009	-0.001
woman	0.010	0.001	0.011	0.002	0.010	0.001	0.008	-0.001	0.006	-0.003	0.011	0.002
mature	0.010	0.001	0.020	0.010	0.007	-0.003	0.006	-0.003	0.001	-0.009	0.014	0.004

Figure 25: Most skewed associations in FineWeb for ‘christian’ compared to other religions, measured using TF-IDF. Columns are sorted by the ‘christian+’ column, measuring the difference from the mean over all words.

word	muslim	muslim+	christian	christian+	jewish	jewish+	hindu	hindu+	buddhist	buddhist+	atheist	atheist+
muslim	0.115	0.083	0.011	-0.021	0.018	-0.015	0.027	-0.006	0.015	-0.017	0.009	-0.024
dating	0.164	0.021	0.192	0.049	0.212	0.069	0.146	0.004	0.133	-0.009	0.009	-0.134
women	0.048	0.007	0.037	-0.004	0.050	0.009	0.053	0.012	0.046	0.006	0.010	-0.030
sex	0.020	0.007	0.015	0.001	0.015	0.002	0.014	0.000	0.011	-0.002	0.005	-0.008
online	0.043	0.006	0.047	0.010	0.057	0.020	0.038	0.001	0.032	-0.005	0.004	-0.033
girl	0.015	0.005	0.009	-0.000	0.011	0.002	0.010	0.001	0.007	-0.002	0.003	-0.006
girls	0.016	0.005	0.012	0.000	0.016	0.004	0.014	0.002	0.010	-0.002	0.003	-0.009
sites	0.019	0.004	0.023	0.009	0.022	0.008	0.013	-0.002	0.009	-0.005	0.002	-0.013
mature	0.014	0.004	0.010	0.001	0.020	0.010	0.007	-0.003	0.006	-0.003	0.001	-0.009
meet	0.028	0.003	0.026	0.001	0.034	0.010	0.027	0.003	0.028	0.004	0.003	-0.021
woman	0.011	0.002	0.010	0.001	0.011	0.002	0.010	0.001	0.008	-0.001	0.006	-0.003
asian	0.011	0.002	0.009	-0.001	0.012	0.002	0.014	0.004	0.011	0.001	0.001	-0.009
free	0.034	0.002	0.038	0.005	0.042	0.009	0.038	0.005	0.035	0.002	0.008	-0.024
site	0.033	0.002	0.035	0.004	0.043	0.012	0.035	0.003	0.038	0.007	0.004	-0.027
looking	0.015	0.002	0.015	0.001	0.018	0.005	0.015	0.001	0.014	0.001	0.004	-0.009
gay	0.013	0.001	0.013	0.001	0.016	0.004	0.013	0.001	0.010	-0.002	0.006	-0.006
best	0.014	0.001	0.015	0.002	0.016	0.003	0.014	0.001	0.013	-0.000	0.006	-0.006
men	0.030	0.001	0.028	-0.002	0.034	0.004	0.040	0.011	0.036	0.007	0.009	-0.021
man	0.012	0.001	0.012	0.001	0.011	0.000	0.009	-0.002	0.007	-0.004	0.015	0.004

Figure 26: Most skewed associations in FineWeb for ‘muslim’ compared to other religions, measured using TF-IDF. Columns are sorted by the ‘muslim+’ column, measuring the difference from the mean over all words.

word	jewish	jewish+	hindu	hindu+	buddhist	buddhist+	atheist	atheist+	muslim	muslim+	christian	christian+
jewish	0.128	0.097	0.018	-0.014	0.013	-0.019	0.007	-0.024	0.012	-0.020	0.012	-0.020
dating	0.212	0.069	0.146	0.004	0.133	-0.009	0.009	-0.134	0.164	0.021	0.192	0.049
singles	0.110	0.041	0.079	0.010	0.085	0.015	0.002	-0.068	0.065	-0.005	0.076	0.006
online	0.057	0.020	0.038	0.001	0.032	-0.005	0.004	-0.033	0.043	0.006	0.047	0.010
site	0.043	0.012	0.035	0.003	0.038	0.007	0.004	-0.027	0.033	0.002	0.035	0.004
mature	0.020	0.010	0.007	-0.003	0.006	-0.003	0.001	-0.009	0.014	0.004	0.010	0.001
meet	0.034	0.010	0.027	0.003	0.028	0.004	0.003	-0.021	0.028	0.003	0.026	0.001
personals	0.032	0.009	0.031	0.009	0.034	0.012	0.001	-0.021	0.018	-0.004	0.018	-0.004
free	0.042	0.009	0.038	0.005	0.035	0.002	0.008	-0.024	0.034	0.002	0.038	0.005
women	0.050	0.009	0.053	0.012	0.046	0.006	0.010	-0.030	0.048	0.007	0.037	-0.004
sites	0.022	0.008	0.013	-0.002	0.009	-0.005	0.002	-0.013	0.019	0.004	0.023	0.009
single	0.045	0.007	0.054	0.017	0.055	0.018	0.003	-0.034	0.034	-0.003	0.033	-0.004
looking	0.018	0.005	0.015	0.001	0.014	0.001	0.004	-0.009	0.015	0.002	0.015	0.001
men	0.034	0.004	0.040	0.011	0.036	0.007	0.009	-0.021	0.030	0.001	0.028	-0.002
girls	0.016	0.004	0.014	0.002	0.010	-0.002	0.003	-0.009	0.016	0.005	0.012	0.000
gay	0.016	0.004	0.013	0.001	0.010	-0.002	0.006	-0.006	0.013	0.001	0.013	0.001
best	0.016	0.003	0.014	0.001	0.013	-0.000	0.006	-0.006	0.014	0.001	0.015	0.002
catholic	0.014	0.003	0.006	-0.004	0.010	-0.001	0.012	0.002	0.006	-0.004	0.015	0.005
new	0.016	0.002	0.013	-0.001	0.013	-0.001	0.014	-0.000	0.013	-0.001	0.014	0.000
date	0.012	0.002	0.012	0.002	0.011	0.002	0.002	-0.008	0.010	0.000	0.012	0.002
asian	0.012	0.002	0.014	0.004	0.011	0.001	0.001	-0.009	0.011	0.002	0.009	-0.001
girl	0.011	0.002	0.010	0.001	0.007	-0.002	0.003	-0.006	0.015	0.005	0.009	-0.000
sex	0.015	0.002	0.014	0.000	0.011	-0.002	0.005	-0.008	0.020	0.007	0.015	0.001
chat	0.016	0.002	0.017	0.003	0.019	0.005	0.001	-0.013	0.014	0.000	0.016	0.002
100	0.011	0.002	0.011	0.002	0.013	0.003	0.002	-0.007	0.008	-0.001	0.010	0.000
woman	0.011	0.002	0.010	0.001	0.008	-0.001	0.006	-0.003	0.011	0.002	0.010	0.001
essay	0.010	0.001	0.017	0.007	0.013	0.003	0.003	-0.007	0.007	-0.003	0.009	-0.001

Figure 27: Most skewed associations in FineWeb for ‘jewish’ compared to other religions, measured using TF-IDF. Columns are sorted by the ‘jewish+’ column, measuring the difference from the mean over all words.

word	hindu	hindu+	buddhist	buddhist+	atheist	atheist+	muslim	muslim+	christian	christian+	jewish	jewish+
hindu	0.129	0.100	0.020	-0.009	0.002	-0.027	0.015	-0.015	0.004	-0.026	0.006	-0.023
indian	0.037	0.026	0.008	-0.003	0.001	-0.009	0.011	0.000	0.004	-0.007	0.004	-0.007
single	0.054	0.017	0.055	0.018	0.003	-0.034	0.034	-0.003	0.033	-0.004	0.045	0.007
women	0.053	0.012	0.046	0.006	0.010	-0.030	0.048	0.007	0.037	-0.004	0.050	0.009
men	0.040	0.011	0.036	0.007	0.009	-0.021	0.030	0.001	0.028	-0.002	0.034	0.004
singles	0.079	0.010	0.085	0.015	0.002	-0.068	0.065	-0.005	0.076	0.006	0.110	0.041
ersonals	0.031	0.009	0.034	0.012	0.001	-0.021	0.018	-0.004	0.018	-0.004	0.032	0.009
essay	0.017	0.007	0.013	0.003	0.003	-0.007	0.007	-0.003	0.009	-0.001	0.010	0.001
free	0.038	0.005	0.035	0.002	0.008	-0.024	0.034	0.002	0.038	0.005	0.042	0.009
asian	0.014	0.004	0.011	0.001	0.001	-0.009	0.011	0.002	0.009	-0.001	0.012	0.002
dating	0.146	0.004	0.133	-0.009	0.009	-0.134	0.164	0.021	0.192	0.049	0.212	0.069
chat	0.017	0.003	0.019	0.005	0.001	-0.013	0.014	0.000	0.016	0.002	0.016	0.002
site	0.035	0.003	0.038	0.007	0.004	-0.027	0.033	0.002	0.035	0.004	0.043	0.012
meet	0.027	0.003	0.028	0.004	0.003	-0.021	0.028	0.003	0.026	0.001	0.034	0.010
date	0.012	0.002	0.011	0.002	0.002	-0.008	0.010	0.000	0.012	0.002	0.012	0.002
100	0.011	0.002	0.013	0.003	0.002	-0.007	0.008	-0.001	0.010	0.000	0.011	0.002
girls	0.014	0.002	0.010	-0.002	0.003	-0.009	0.016	0.005	0.012	0.000	0.016	0.004
gay	0.013	0.001	0.010	-0.002	0.006	-0.006	0.013	0.001	0.013	0.001	0.016	0.004
love	0.016	0.001	0.014	-0.001	0.013	-0.002	0.013	-0.001	0.017	0.003	0.014	-0.000
looking	0.015	0.001	0.014	0.001	0.004	-0.009	0.015	0.002	0.015	0.001	0.018	0.005
girl	0.010	0.001	0.007	-0.002	0.003	-0.006	0.015	0.005	0.009	-0.000	0.011	0.002
online	0.038	0.001	0.032	-0.005	0.004	-0.033	0.043	0.006	0.047	0.010	0.057	0.020
best	0.014	0.001	0.013	-0.000	0.006	-0.006	0.014	0.001	0.015	0.002	0.016	0.003
woman	0.010	0.001	0.008	-0.001	0.006	-0.003	0.011	0.002	0.010	0.001	0.011	0.002

Figure 28: Most skewed associations in FineWeb for ‘hindu’ compared to other religions, measured using TF-IDF. Columns are sorted by the ‘hindu+’ column, measuring the difference from the mean over all words.

word	old	old+	young	young+
old	0.034	0.012	0.010	-0.012
just	0.019	0.002	0.016	-0.002
new	0.018	0.002	0.015	-0.002
like	0.021	0.002	0.017	-0.002
ve	0.011	0.001	0.009	-0.001
don	0.013	0.001	0.010	-0.001
ll	0.009	0.001	0.006	-0.001
use	0.009	0.001	0.006	-0.001
time	0.019	0.001	0.017	-0.001
really	0.012	0.001	0.009	-0.001
good	0.013	0.001	0.011	-0.001
little	0.010	0.001	0.008	-0.001
got	0.009	0.001	0.007	-0.001
things	0.010	0.001	0.008	-0.001
know	0.013	0.001	0.011	-0.001
make	0.012	0.001	0.011	-0.001
want	0.010	0.001	0.009	-0.001
look	0.008	0.001	0.007	-0.001
need	0.009	0.001	0.008	-0.001
home	0.010	0.001	0.009	-0.001
right	0.009	0.001	0.008	-0.001
going	0.010	0.001	0.009	-0.001
day	0.012	0.001	0.011	-0.001
way	0.012	0.001	0.011	-0.001
great	0.010	0.001	0.009	-0.001
think	0.011	0.001	0.010	-0.001

A

word	young	young+	old	old+
young	0.038	0.016	0.006	-0.016
children	0.016	0.005	0.007	-0.005
women	0.011	0.003	0.006	-0.003
school	0.013	0.003	0.008	-0.003
said	0.017	0.002	0.012	-0.002
people	0.021	0.002	0.016	-0.002
child	0.009	0.002	0.005	-0.002
life	0.015	0.001	0.012	-0.001
family	0.011	0.001	0.008	-0.001
story	0.009	0.001	0.007	-0.001
world	0.012	0.001	0.010	-0.001
man	0.010	0.001	0.008	-0.001
book	0.010	0.001	0.008	-0.001

B

Figure 29: *Age* bias in FineWeb, measured as most skewed associations for 'old' and 'young', using TF-IDF. Sorted by the difference from the mean TF-IDF for all words associated to 'old' ('old+', A) and 'young' ('young+', B).