Dataset	Transparency		Versatility			Scale (TB)
	Open Access	Open Code	Raw Data	Composite	Multilingual	
Refined Web	✔(subset)	✗	✗	✗	✗	2.8
FineWeb	✔	✔	✗	✗	✗	93.4
FineWeb EDU	✔	✔	✗	✗	✗	8.8
C4	✔	✔	✗	✗	✗	0.3
mC4	✔	✔	✗	✗	✔	9.7
DCLM baseline	✔	✔	✗	✗	✗	10.0
DCLM-Pool	✔	✔	✔	✗	✔	340.0
Dolma v1.7	✔	✔	✗	✔	✗	4.5
Pile	✔	✔	✗	✔	✗	0.8
SlimPajama	✔	✔	✗	✔	✗	0.9
ROOTS	✔	✔	✗	✔	✔	1.6
RedPajama-V1	✔	✔	✗	✔	✗	3.0
RedPajama-V2	✔	✔	✔	✗	✔	270.0

Table 1: Comparison of open Pretrainging Datasets along the dimensions of transparency, versatility, and scale.

Subset	Uncertainty	Decision
CommonCrawl	Which snapshots were used?	We use the first snapshot from 2019 to 2023.
	What classifier was used, and how was it constructed?	We use a fasttext classifier with unigram features and use 300k training samples.
	What threshold was used to classify a sample as high quality?	We set the threshold to match the token count reported in LLama.
GitHub	Quality filtering heuristics	We remove any file • with a maximum line length of more than 1000 characters. • with an average line length of more than 100 characters. • with a proportion of alphanumeric characters of less than 0.25. • with a ratio between the number of alphabetical characters and the number of tokens of less than 1.5. • whose extension is not in the following set of whitelisted extensions: .asm, .bat, .cmd, .c, .h, .cs, .cpp, .hpp, .c++, .h++, .cc, .hh, .C, .H, .cmake, .css, .dockerfile, .f90, .f, .f03, .f08, .f77, .f95, .for, .fpp, .go, .hs, .html, .java, .js, .jl, .lua, .md, .markdown, .php, .php3, .php4, .php5, .phps, .phpt, .pl, .pm, .pod, .perl, .ps1, .psd1, .psm1, .py, .rb, .rs, .sql, .scala, .sh, .bash, .command, .zsh, .ts, .tsx, .tex, .vb, Dockerfile, Makefile, .xml, .rst, .m, .smali
Wikipedia	Which Wikipedia dump was used?	We used the most recent at the time of data curation (2023-03-20).
Books	How were the books deduplicated?	We use SimHash to perform near deduplication.

Table 2: Overview over the different uncertainties and decisions made during the construction of the RedPajama-V1 dataset.