Source	Type	Tokens	Words	Bytes	Docs
Pretraining ✦ OLMo 2 1124 Mix					
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

Table 1 Composition of the pretraining data for OLMo 2. The OLMo 2 1124 Mix is composed of StarCoder (Li et al., 2023b; Kocetkov et al., 2022), peS2o (Soldaini and Lo, 2023), web text from DCLM (Li et al., 2024) and Wiki come from Dolma 1.7 (Soldaini et al., 2024). arXiv comes from Red-Pajama (Together AI, 2023), while OpenWebMath (Paster et al., 2023) and Algebraic Stack come from ProofPile II (Azerbayev et al., 2023).

2.1.1 Pretraining data: OLMo 2 Mix 1124

The mix used for this stage is shown in Table 1. It consists of approximately 3.9 trillion tokens, with over 95% derived from web data. We refer to this set as OLMo 2 Mix 1124. This is the same pretraining data used in OLMoE (Muennighoff et al., 2024).

We combine data from DCLM (Li et al., 2024) and Dolma 1.7 (Soldaini et al., 2024). From DCLM, we use the "*baseline 1.0*" mix.[4] From Dolma, we use the arXiv (Together AI, 2023), OpenWebMath (Paster et al., 2023), Algebraic Stack, peS2o (Soldaini and Lo, 2023), and Wikipedia subsets. arXiv, OpenWebMath, and Algebraic Stack were originally part of ProofPile II (Azerbayev et al., 2023).

Finally, we include code from StarCoder (Li et al., 2023b), which is derived from permissively-licensed repositories from GitHub (Kocetkov et al., 2022). In an attempt to include higher quality code, we remove any document from a repository with fewer than 2 stars on GitHub. Further, through manual inspection of this source, we found it to contain documents encoded in binary format or containing mostly numerical content; to remove them, we discarded documents whose most frequent word constitutes over 30% of the document, or whose top-2 most frequent words constitute over 50% of the document. To mitigate possible training loss spikes, we remove documents with repeated sequences of 32 or more n-grams. We report details and show effectiveness of this intervention in Section §3.1.

2.1.2 Mid-training data: Dolmino Mix 1124

After the initial pretraining stage on mostly web data, we further train with a mixture of web data that has been more restrictively filtered for quality and a collection of domain-specific high quality data, much of which is synthetic. The purpose of this mixture is to imbue the model with math-centric skills and provide focused exposure to STEM references and high quality text. We generate several variants of this mixture, with varying sizes, but generally refer to this mixture as Dolmino Mix 1124. The base sources from which Dolmino Mix 1124 is subsampled are described in Table 2. We refer the reader to Section §4 for a **deep dive** detailing our processes for experimenting and curating data for this mix.