		Pre-training		Mid-training		Decay Phase	
Category	**Dataset**	**Tokens (B)**	**%**	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Scientific	Arxiv	28.0	1.6	4.1	1.6	3.0	3.9
Reference	Books	5.3	0.3	0.8	0.3	10.5	13.8
Crawl	CC Head	356.6	20.9	–	–	–	–
News	CC News	7.3	0.4	–	–	–	–
Code	Code_Repos	–	–	–	–	20.2	26.5
Crawl	DCLM	837.2	49.1	–	–	–	–
Crawl	DCLM (Dolmino)	–	–	175.5	70.4	26.0	34.1
Math	Math (Dolmino)	–	–	10.4	4.2	5.0	6.6
Math	Open-Web-Math	12.7	0.7	–	–	–	–
Math	Algebraic StackExchange	12.6	0.7	–	–	–	–
Scientific	PeS2o	57.3	3.4	8.3	3.3	–	–
Social	Reddit	80.3	4.7	6.2	2.5	–	–
Social	StackExchange	19.6	1.1	–	–	–	–
Social	StackExchange (Dolmino)	–	–	2.7	1.1	4.0	5.2
Code	Starcoder	263.9	15.5	38.4	15.4	–	–
Reference	Textbooks	–	–	–	–	0.5	0.7
Instruction	Tulu Flan	16.6	1.0	2.4	1.0	4.1	5.4
Reference	Wikipedia	7.3	0.4	0.5	0.2	3.0	3.9
Total		1,704.7	100.0	249.3	100.0	76.3	100.0