

Midtraining Data Sources

Dolmino 1 Flan	2e4	6e3	0	0	0	809	124	0	127	10	0	0	35	0	0	6e3	68	3e3	0	0	0	0	0	0
Tulu 3 SFT	1e4	270	0	14	0	85	42	2	25	5	0	0	3	0	0	287	6e3	893	554	1e3	1e3	689	260	14
Nemotron Synth QA	6e3	692	10	2e3	79	2	23	167	0	80	74	17	15	31	22	519	689	27	1e3	41	64	189	2	0
Dolmino Math	4e3	0	2e3	49	1e3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0
Common Crawl (High Q.)	2e3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2e3	50	0	256	0	1	119	0	0
StackEdu (FIM)	876	0	0	0	0	0	0	0	0	0	0	0	0	0	0	792	0	0	24	0	1	58	0	1
Gemini Reasoning Traces	606	0	513	31	0	0	0	0	0	0	0	43	0	19	0	0	0	0	0	0	0	0	0	0
OLMOOCR Science PDFs (High Q.)	554	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	4	1	390	0	19	33	10	0
Sponge	308	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	1	0	190	0	0	79	0	0
General Reasoning Mix	113	0	6	68	3	0	0	0	0	0	0	5	0	0	0	0	0	0	4	0	27	0	0	0

Occurrences Of Contamination

Total
contam

Evaluated splits:

Val/Test

All

Perf Δ	1.7	2.0	-1.2	-1.6	13.9	0.4	-0.4	-2.4	0.6	-0.1	-0.7	-0.0	0.6	0.9	-1.4	0.0	1.4	-0.3	1.8	1.1	-0.4	1.1	-0.3
% contam	27%	50%	4%	100%	9%	2%	2%	2%	2%	0%	5%	1%	24%	3%	6%	3%	2%	13%	3%	2%	6%	0%	0%

SQuAD	Minerva	MMLU (MC)	GSM8K	DROP	CoQA (MC)	HumEval (@16)	DROP (MC)	LAMBADA	MedMCQA (MC)	MedQA En (MC)	SQuAD (MC)	LeetCode (@16)	M-E-HumEval (@16)	Jeopardy	HellaSwag	CoQA	ARC (MC)	PIQA (MC)	CSQA (MC)	SciQ (MC)	Winogrande	SocialIQA (MC)
-------	---------	-----------	-------	------	-----------	---------------	-----------	---------	--------------	---------------	------------	----------------	-------------------	----------	-----------	------	----------	-----------	-----------	-----------	------------	----------------

Benchmark (Metric)