

C Additional Information

We present additional information for the sake of space in the main paper.

Filtered task names. We present task names we use to filter FLAN derived datasets such as OpenOrca in Table 8.

Filtered Task Name
task228_arc_answer_generation_easy
ai2_arcARCCChallenge:1.0.0
ai2_arcARCEasy:1.0.0
task229_arc_answer_generation_hard
hellaswag:1.1.0
task1389_hellaswag_completion
cot_gsm8k
cot_gsm8k_ii
drop:2.0.0
winogrande:1.1.0

Table 8: Task names that we use to filter data for FLAN derived datasets such as OpenOrca.

ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
0.06	N/A	0.15	0.28	N/A	0.70

Table 9: Data contamination test results for SOLAR 10.7B-Instruct. We show ‘result < 0.1, %’ values where a value higher than 0.9 indicates high probability of data contamination. HellaSwag and Winogrande datasets are not currently supported. We set SOLAR 10.7B as our reference model when performing the data contamination tests.

Results on data contamination. To show the integrity of SOLAR 10.7B-Instruct, we also report the data contamination test (Shi et al., 2023) results in Table. 9. All four tested benchmark datasets yield results well below the contamination threshold, affirming the absence of data contamination in our model. One interesting point is that the value for GSM8K is noticeably higher than for other datasets, even without contamination. One potential reason for this is the stronger data similarity in math-related instruction datasets.