Model	Size	Type	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
SOLAR 10.7B-Instruct	$\sim$ 11B	Alignment-tuned	**74.20**	**71.08**	88.16	66.21	**71.43**	83.58	64.75
Qwen 72B	$\sim$ 72B	Pretrained	73.60	65.19	85.94	**77.37**	60.19	82.48	**70.43**
Mixtral 8x7B-Instruct-v0.1	$\sim$ 47B	Instruction-tuned	72.62	70.22	87.63	71.16	64.58	81.37	60.73
Yi 34B-200K	$\sim$ 34B	Pretrained	70.81	65.36	85.58	76.06	53.64	82.56	61.64
Yi 34B	$\sim$ 34B	Pretrained	69.42	64.59	85.69	76.35	56.23	83.03	50.64
Mixtral 8x7B-v0.1	$\sim$ 47B	Pretrained	68.42	66.04	86.49	71.82	46.78	81.93	57.47
Llama 2 70B	$\sim$ 70B	Pretrained	67.87	67.32	87.33	69.83	44.92	83.74	54.06
Falcon 180B	$\sim$ 180B	Pretrained	67.85	69.45	**88.86**	70.50	45.47	**86.90**	45.94
SOLAR 10.7B	$\sim$ 11B	Pretrained	66.04	61.95	84.60	65.48	45.04	83.66	55.50
Qwen 14B	$\sim$ 14B	Pretrained	65.86	58.28	83.99	67.70	49.43	76.80	58.98
Mistral 7B-Instruct-v0.2	$\sim$ 7B	Instruction-tuned	65.71	63.14	84.88	60.78	68.26	77.19	40.03
Yi 34B-Chat	$\sim$ 34B	Instruction-tuned	65.32	65.44	84.16	74.90	55.37	80.11	31.92
Mistral 7B	$\sim$ 7B	Pretrained	60.97	59.98	83.31	64.16	42.15	78.37	37.83

Table 2: Evaluation results in the Open LLM Leaderboard for SOLAR 10.7B and SOLAR 10.7B-Instruct along with other top-performing models. We report the scores for the six tasks mentioned in Sec. 4.1 along with the H6 score (average of six tasks). We also report the size of the models in units of billions of parameters. The type indicates the training stage of the model and is chosen from {Pretrained, Instruction-tuned, Alignment-tuned}. Models based on SOLAR 10.7B are colored purple. The best scores for H6 and the individual tasks are shown in bold.

MetaMathQA (Yu et al., 2023) dataset.

We reformatted the instruction datasets with an Alpaca-styled chat template. For datasets such as OpenOrca, which are derived from FLAN (Longpre et al., 2023), we filter data that overlaps with the benchmark datasets (see Tab. 8 in Appendix. C for more information). The alignment datasets are in the {prompt, chosen, rejected} triplet format. We preprocess the alignment datasets following Zephyr (Tunstall et al., 2023). We use Dataverse (Park et al., 2024) for data preprocessing.

Evaluation. In the HuggingFace Open LLM Leaderboard (Beeching et al., 2023), six types of evaluation methods are presented: ARC (Clark et al., 2018), HellaSWAG (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2022), Winogrande (Sakaguchi et al., 2021), and GSM8K (Cobbe et al., 2021). We utilize these datasets as benchmarks for evaluation and also report the average scores for the six tasks, *e.g.,* H6. We either submit directly to the Open LLM Leaderboard or utilize Evalverse (Kim et al., 2024b) for running evaluations locally.

Model merging. Model merging methods such as Yadav et al. (2023) can boost model performance without further training. We merge some of the models that we trained in both the instruction and alignment tuning stages. We implement our own merging methods although popular open source also exist such as MergeKit[3].

4.2 Main Results

We present evaluation results for our SOLAR 10.7B and SOLAR 10.7B-Instruct models along with other top-performing models in Tab. 2. SOLAR 10.7B outperforms other pretrained models of similar sizes, such as Qwen 14B and Mistral 7B, which shows that DUS is an effective method to up-scale base LLMs. Furthermore, despite the smaller size, SOLAR 10.7B-Instruct scores the highest in terms of H6, even surpassing the recent top-performing open-source LLM Mixtral 8x7B-Instruct-v0.1 or Qwen 72B. The above results indicate DUS can up-scale models that are capable of achieving state-of-the-art performance when finetuned. We also report data contamination results for SOLAR 10.7B-Instruct in Appendix C.

4.3 Ablation Studies

We present ablation studies for both the instruction and alignment tuning stages. Note that the evaluation results for the following studies are ran locally and may vary from results obtained by submitting to the Open LLM Leaderboard.

4.3.1 Instruction Tuning

Ablation on the training datasets. We present ablation studies using different training datasets for the instruction tuning in Tab. 3. The ablated models are prefixed with SFT for supervised finetuning. 'SFT v1' only uses the Alpaca-GPT4 dataset, whereas 'SFT v2' also uses the OpenOrca dataset. 'SFT v3' uses the Synth. Math-Instruct dataset along with the datasets used in 'SFT v2'. Similarly, 'SFT v4' uses the Synth. Math-Instruct dataset along with the datasets used in 'SFT v1'.

First, we analyze how Alpaca-GPT4 and OpenOrca affect the trained models. The first ablated model, 'SFT v1', which used only the Alpaca-GPT4 dataset for training, resulted in 69.15 for H6.

[3] https://github.com/cg123/mergekit