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

Model	Alpaca-GPT4	OpenOrca	Synth. Math-Instruct	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
SFT v1	O	✗	✗	69.15	**67.66**	**86.03**	65.88	**60.12**	**82.95**	52.24
SFT v2	O	O	✗	69.21	65.36	85.39	65.93	58.47	82.79	57.32
SFT v3	O	O	O	70.03	65.87	85.55	65.31	57.93	81.37	64.14
SFT v4	O	✗	O	70.88	67.32	85.87	65.87	58.97	82.48	64.75
SFT v3 + v4	O	O	O	**71.11**	67.32	85.96	**65.95**	58.80	82.08	**66.57**

Table 3: Ablation studies on the different datasets used for instruction tuning. 'SFT v3+v4' indicates that the model is merged from 'SFT v3' and 'SFT v4' by simply averaging the model weights. The best scores for H6 and the individual tasks are shown in bold.

Model	Ultrafeedback Clean	Synth. Math-Alignment	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
DPO v1	O	✗	73.06	71.42	**88.49**	**66.14**	72.04	81.45	58.83
DPO v2	O	O	**73.42**	**71.50**	88.28	65.97	71.71	**82.79**	**60.27**
DPO v1 + v2	O	O	73.21	71.33	88.36	65.92	**72.65**	**82.79**	58.23

Table 4: Ablation studies on the different datasets used during the direct preference optimization (DPO) stage. 'SFT v3' is used as the SFT base model for DPO. We name ablated models with the 'DPO' prefix to indicate the alignment tuning stage. 'DPO v1+v2' indicates that the model is merged from 'DPO v1' and 'DPO v2' by simply averaging the model weights. The best scores for H6 and the individual tasks are shown in bold.

Model	Base SFT Model	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
DPO v2	SFT v3	73.42	**71.50**	**88.28**	**65.97**	71.71	**82.79**	60.27
DPO v3	SFT v3 + v4	**73.58**	71.33	88.08	65.39	**72.45**	81.93	**62.32**

Table 5: Ablation studies on the different SFT base models used during the direct preference optimization (DPO) stage. Ultrafeedback Clean and Synth. Math-Alignment datasets are used. We name ablated models with the 'DPO' prefix to indicate the alignment tuning stage. The best scores for H6 and the individual tasks are shown in bold.

When we add the OpenOrca dataset to train the second ablated model, 'SFT v2', the resulting H6 score is 69.21, which is little change from 69.15 of 'SFT v1'. However, the task scores vary more as 'SFT v2' gets a substantially higher GSM8K score of 57.32 compared to 52.24 of 'SFT v1' but also gets noticeably lower scores across the board for ARC, HellaSwag, and TruthfulQA. This seems to indicate that using OpenOrca results in a model that behaves differently from using only Alpaca-GPT4.

Second, we investigate whether Synth. Math-Instruct dataset is beneficial. For 'SFT v3', we add the Synth. Math-Instruct dataset, which boosts GSM8K scores to 64.14 and achieves comparable scores for the other tasks. Interestingly, when we add the Synth. Math-Instruct dataset to 'SFT v1' to train 'SFT v4', we get our highest H6 score of 70.88 with higher scores than 'SFT v3' for all tasks. From the above, we can see that adding the Synth. Math-Instruct dataset is helpful.

Lastly, we see whether merging models trained with and without OpenOrca can boost performance. In the first analysis, we saw that using OpenOrca resulted in a model that behaved differently from the model that was trained without OpenOrca. Building on this intuition, we merge 'SFT v3' and 'SFT v4' as they are the best-performing models with and without OpenOrca. To our surprise, the resulting merged model 'SFT v3+v4' retains the high scores for non-GSM8K tasks from 'SFT v4' but also achieves a higher GSM8K score than 'SFT v3' or 'SFT v4'. Thus, we see that merging models that specialize in different tasks is a promising way to obtain a model that performs well generally.

4.3.2 Alignment Tuning

As we utilize sDPO for practical alignment tuning, there are additional aspects to ablate such as the SFT base models used. Thus, we present ablations for the different training datasets used for training, the different SFT base models to initialize the sDPO training, and finally, the model merging strategy to obtain the final alignment-tuned model.

Ablation on the training datasets. We ablate on the different alignment datasets used during DPO in Tab. 4. We use 'SFT v3' as the SFT base model for DPO. 'DPO v1' only uses the Ultrafeedback Clean dataset while 'DPO v2' also used the Synth. Math-Alignment dataset.

First, we test how Ultrafeedback Clean and Synth. Math-Alignment impacts model performance. For 'DPO v1', it achieves 73.06 in H6, which is a substantial boost from the SFT base model score of 70.03. However, we note that while