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

Model	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
Cand. 1	**73.73**	70.48	87.47	65.73	70.62	81.53	**66.57**
Cand. 2	73.28	**71.59**	**88.39**	**66.14**	**72.50**	**81.99**	59.14

Table 6: Performance comparison amongst the merge candidates. 'Cand. 1' and 'Cand. 2' are trained using the same setting as 'DPO v2' and 'DPO v3', respectively, but with slightly different hyper-parameters. The best scores for H6 and the individual tasks are shown in bold.

Model	Merge Method	H6 (Avg.)	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
Merge v1	Average (0.5, 0.5)	74.00	**71.16**	88.01	66.14	71.71	**82.08**	64.90
Merge v2	Average (0.4, 0.6)	73.93	71.08	**88.08**	**66.27**	**71.89**	81.77	64.52
Merge v3	Average (0.6, 0.4)	**74.05**	71.08	87.88	66.13	71.61	**82.08**	**65.50**
Merge v4	SLERP	73.96	**71.16**	88.03	66.25	71.79	81.93	64.59

Table 7: Ablation studies on the different merge methods used for obtaining the final model. We use 'Cand. 1' and 'Cand. 2' from Tab. 6 as our two models for merging. We name the merged models with the 'Merge' prefix to indicate they are merged. The best scores for H6 and the individual tasks are shown in bold.

scores for tasks like ARC, HellaSwag, and TruthfulQA all improved by good margins, the score for GSM8K is 58.83, which is lower than the SFT base model score of 64.14. Adding Synth. Math-Alignment to train 'DPO v2', we see that the GSM8k score improves to 60.27, which is lower than the SFT base model but still higher than 'DPO v1'. Other task scores are also not negatively impacted by adding Synth. Math-Alignment. Thus, we can conclude that adding Synth. Math-Alignment is beneficial for H6.

Then, we experiment whether merging 'DPO v1' and 'DPO v2' is beneficial. Unfortunately, 'DPO v1+v2' scores 73.21 in H6, which is worse than 'DPO v2'. More importantly, the gain in the GSM8K score from adding Synth. Math-Alignment is gone, which is undesirable. One reason for this could be that 'DPO v2' is a strict improvement over 'DPO v1', unlike the case for merging 'SFT v3' and 'SFT v4' where the models had different strengths and weaknesses.

Ablation on the SFT base models. When applying DPO, we start from a model that is already instruction tuned ,*i.e.,* the SFT base model and ablate on using different SFT base models. We use Ultrafeedback Clean and Synth. Math-Alignment datasets for this ablation. Each of the ablated models is trained as follows. 'DPO v2' uses 'SFT v3' as the base SFT model, while 'DPO v3' uses 'SFT v3+v4' as the SFT base model instead.

Note that 'SFT v3+v4' has higher scores on all tasks compared to 'SFT v3', and the gap is especially large for ARC ($+1.45$) and GSM8K ($+2.43$). Surprisingly, the two models perform similarly in terms of H6. A closer look at the scores for the individual tasks shows only a small margin in the GSM8K scores, and other task scores show little difference. Thus, the performance gaps in certain tasks in the SFT base models do not always carry over to the alignment-tuned models.

Ablation on different merge methods. From Tab. 3, we saw that merging two models that have different strengths can be beneficial to performance. To utilize this for the alignment-tuned model as well, we train two models named 'Cand. 1' and 'Cand. 2' using the same training dataset and SFT base model as 'DPO v2' and 'DPO v3' but with different hyper-parameters to maximize each model's respective strengths. We compare 'Cand. 1' and 'Cand. 2' in Tab. 6 where we can see that 'Cand. 1' has high GSM8K scores but relatively low scores for the other tasks, whereas 'Cand. 2' has low scores for GSM8K but high scores for the other tasks. We merge these two models using various methods and ablate the results in Tab.. 7.

We use two merge methods: 1) Average (a, b), where a and b denote the weighting for 'Cand. 1' and 'Cand. 2' when averaging weights and 2) SLERP (Shoemake, 1985). We use (0.5, 0.5), (0.4, 0.6), and (0.6, 0.4) for Average (a, b). From Tab. 7, we can see that the different merge methods have little effect on the H6 scores. The scores for the individual tasks also do not differ by much, suggesting that as long as the merge candidates have sufficiently different strengths, the exact merge method may not be as crucial. Thus, we chose 'Merge v1' as our SOLAR 10.7B-Instruct model.