Dataset	Editor		
	Original model	Original model+Forgetting	Original model+F-learning
GSM8K	2.35	2.5	1.44
MATH	3.2	3	1.96
MMLU-college-chemistry	24	34	41
MMLU-college-mathematics	29	30	26
MMLU-management	16.50	14.56	32.04
MMLU-computer-security	21	20	23
MMLU-macroeconomics	32.31	33.08	34.87
MMLU-college-physics	19.61	26.47	22.55
MMLU-astronomy	30.92	23.03	32.24
MMLU-professional-law	26.47	25.62	24.51
MMLU-college-medicine	24.28	24.28	29.48

Table 5: Results on accuracy of the GSM8K, MATH and MMLU dataset based on LLAMA2-7B.

and gradient-accumulation-steps is 4.

When we used the Deepspeed, we set 4 processes and zero-stage is 2.

A.2.2 Full-FT and LoRA

Full-FT and LoRA refer to knowledge updating by full fine-tuning and LoRA fine-tuning in our experiments. We adopted experimental settings similar to F-learning as mentioned above. The difference is that these two do not forget the old knowledge. Full-FT and LoRA also use the instruction fine-tuning data mentioned in for supervised fine-tuning training. Instruction fine-tuning can make it generate answers to prompts better.

A.2.3 FT-c

Knowledge updating of FT-c is executed at layer 21, where optimization proceeds for 5 steps with a learning rate of 5e-5. And the batch-size is 1.

A.2.4 ROME

Knowledge updating of ROME is executed at layer 5, where optimization proceeds for 25 steps with a learning rate of 5e-3. And the weight-decay is 1e-3, the kl-factor is 0.0625. Covariance statistics are collected in float32 on Wikitext using a sample size of 100,000.

A.2.5 MEMIT

Knowledge updating of ROME is executed at layer n = [4, 5, 6, 7, 8], where optimization proceeds for 25 steps with a learning rate of 5e-2. The batch is the 19,085 (or 10,000). And the weight-decay is 1e-3, the kl-factor is 0.0625. Covariance statistics are collected in float32 on Wikitext using a sample size of 100,000.

A.3 Impact to Other Capabilities within LLMs Testing

Here we test the impact of forgetting old knowledge and learning new knowledge over the original model on other capabilities of the model (such as mathematical abilities). Specifically, we evaluated changes in the model's mathematical capabilities on GSM8K and MATH, and evaluated changes in the model's comprehensive examining capabilities on MMLU. Taking LLAMA2-7B as an example. The results are shown in the Table 5. "Original model+Forgetting" refers to only forgetting the old knowledge over the original model, and "Original model+F-learning" refers to the original model with our F-learning method. The hyper-parameters λ of the rate of forgetting is set to 0.3. And the metric of evaluation in experiments is mainly "Accuracy". Experiments show that F-learning can slightly improve the mathematical ability of the model after forgetting old knowledge, but it will decrease after learning new knowledge. While in other areas of professional capabilities, our method has little impact. Interestingly, the model's performance on "college-chemistry" and "college-medicine" has been significantly improved after completing the knowledge update. This may be because the dataset contains relevant knowledge.

A.4 Interpretability of Parametric Arithmetic

Recently, Parametric Arithmetic has become a common method for parameter fine-tuning because of its operability and adaptability. previous work (Ilharco et al., 2022) has conducted experimental research on parameter parametric arithmetic and verified that subtracting the parameters obtained by