part of the old knowledge during the knowledge updating process. Given that our original model has fully learned a large quantity of old knowledge, it faces greater challenges in learning new knowledge. Surprisingly, ROME maintains Reliability and Generality almost unchanged from the original model on two datasets, while having a high locality (more than 90), which suggests that it performs little knowledge updating (As the injection of new knowledge will usually have an impact on locality), likely because that it can only edit a small number of parameters, coupled with the fact that our original model is full of old knowledge, which obstructs the effect of the causal tracing mechanism in ROME. It is worth noting that full fine-tuning is much more capable of learning new knowledge than LoRA, as LoRA focuses on training a limited subset of parameters within the attention structure, while the majority of factual knowledge is encoded in the MLP layers.

5.5 Forgetting with LoRA and Learning with Full Fine-tuning

In the above setting of experiments, the method we adopt is to perform old knowledge forgetting and new knowledge learning based on full (or LoRA) fine-tuning at the same time. Nonetheless, we find that subtracting the knowledge parameters of full fine-tuning (i.e. forgetting old knowledge by full fine-tuning) in some cases will completely destroy the core functions of our base model, resulting in a significant drop in evaluation metrics. In this view, as LoRA is a parameter-efficient fine-tuning method that has less impact on parameters compared to full fine-tuning, we try a new method that forgets old knowledge by LoRA and then learns new knowledge by full fine-tuning as a trade-off. Similar to the method above (§4), we define this process as follows:

$$\theta_{\triangle}^{old} = \mathrm{LoRA}\{\theta, \mathrm{K_{old}}\} - \theta, \tag{6}$$

$$\theta' = \theta - \lambda\theta_{\triangle}^{old}, \tag{7}$$

$$\theta^* = \mathrm{FT}\{\theta', \mathrm{K_{new}}\} \tag{8}$$

where LoRA is the operation of supervised fine-tuning by LoRA while FT is the operation of supervised fine-tuning by full fine-tuning. θ^* is noted as the parameters of the edited model f_{θ^*} which has completed the knowledge updating.

For verification, we keep the same experimental settings as above and conduct the experiment.

Editor	Metric		
	Reliability	Generality	Locality
Original model	27.99	27.89	/
LoRA	29.25	29.07	77.17
F-Learning$_{\mathrm{LoRA}}$	**29.27**	**29.11**	**77.40**
Full-FT	44.60	43.52	63.74
F-Learning$_{\mathrm{LoRA-FT}}$	44.70	43.71	65.50
F-Learning$_{\mathrm{FT}}$	**44.79**	**43.83**	**69.26**

Table 2: Results on three metrics of the zsRE dataset based on BLOOM-7B.

The results are shown in Table 1. The results support that the method of forgetting with LoRA and then learning with full fine-tuning (noted as **F-Learning**$_{\mathrm{LoRA-FT}}$) completely surpassing many baselines such as the directly full fine-tuning, as well as almost nearly match or even surpassing the method of forgetting and learning with full fine-tuning in some cases. In particular, it generally maintains the highest results in locality, which may be owing to the parameter-efficiency of LoRA-based old knowledge forgetting.

After conducting experiments, we empirically discovered that utilizing the technique of forgetting through the subtraction of LoRA parameters can approximate the effect achieved by subtracting the parameters during full fine-tuning. We hold that although LoRA-based knowledge forgetting does not eliminate the old knowledge stored in the MLP layers of the LLMs, it alters the patterns and relationships associated with the old knowledge stored in the attention structure (i.e., an implicit knowledge representation), which facilitates the new knowledge learning. This finding holds significant value due to the considerable reduction in time and computational costs associated with LoRA compared to full fine-tuning.

5.6 Adaptability Testing

To further verify the adaptability of the method, we conducted experiments on zsRE based on BLOOM-7B and maintained the same experimental settings as the above. The results are shown in Table 2. We could find that F-learning still performs well. Notably, although the Reliability and Generality remain roughly stable, the locality is significantly improved, which means that our method could inject new knowledge into the LLM with less cost (As changes to model parameters will inevitably affect the model's locality), demonstrating the necessity and effectiveness of forgetting old knowledge.