fine-tuning can achieve a decrease in indicators on a dataset. On the contrary, adding parameters obtained by fine-tuning can endow the model with capabilities or achieve multi-task learning. Therefore, we believe that by subtracting the parameters trained on old knowledge, it may forget old knowledge and further achieve better knowledge updating. Our experimental results also support this conclusion.

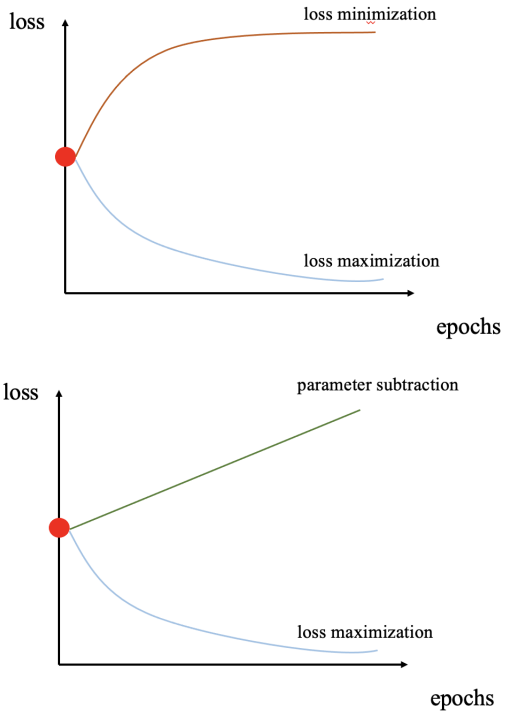

Figure 3: Loss changes of loss maximization and parameter subtraction.

The parametric arithmetic in this paper mainly focuses on parameter subtraction. We subtract parameters so that the loss function moves in the opposite direction to the direction of gradient descent to accumulate errors. This is essentially similar to gradient ascent and loss maximization, a method commonly used in machine unlearning, through the maximum of loss to accumulate errors and damage the performance of the model on a dataset. Their relationship is roughly shown in the figure 3. We do not directly use loss maximization because compared with it, the method of subtracting parameters is more stable and controllable, which can avoid affecting other irrelevant knowledge as much as possible.

A.5 Prospects and Application Scenarios

In an era when LLM's research and application are becoming more and more popular, knowledge update is gaining its attention as a technology for updating the internal knowledge within the LLM. Knowledge updating is closely related to some research fields such as continual learning and machine unlearning. The purpose of knowledge updating is to correct old or wrong knowledge, while continual learning hopes to not forget old knowledge while the LLM continues to learn new knowledge. The aim of machine unlearning is to let the LLM forget harmful or wrong knowledge.

We believe that our method of old knowledge forgetting has a wide range of application scenarios, such as harmful knowledge forgetting, copyright content elimination, user privacy protection for LLMs, etc.