edge that "the Earth is round" when they become adults. Conversely, if they can forget the erroneous knowledge that "the Earth is flat" or if they learn the correct knowledge that "the Earth is round" before being exposed to the incorrect information, it would be much simpler.

Inspired by the above empirical observations and (Ilharco et al., 2022)'s task arithmetic, we propose a novel paradigm of knowledge updating called **F-Learning** (Forgetting before Learning). Specifically, we first fine-tune the initial model using old knowledge and then subtract the difference between the fine-tuned model parameters and the initial model parameters from the initial model parameters. This process is defined as "*old knowledge forgetting*". We then use the new knowledge to fine-tune the model after forgetting the old knowledge. This process we define as "*new knowledge learning*". After the two stages of *forgetting old knowledge* and *learning new knowledge*, the model's knowledge is updated. The contribution of this work can be summarised as follows:

- We propose a novel fine-tuning paradigm "Forgetting before Learning" (F-Learning) for knowledge updating in large language models.
- Experimental results show that our proposed F-Learning improves the knowledge updating performance of various fine-tuning methods and outperforms the existing baselines in most cases.
- Experimental results show that forgetting by subtracting the parameters of LoRA can achieve the approximate effect of subtracting the parameters of full fine-tuning.

2 Related Work

Currently, the method of knowledge updating and model editing (also known as knowledge editing) for LLMs is mainly divided into two classes (Yao et al., 2023; Wang et al., 2023):

a. The method preserving model's parameters

Retrieve augmentation practically depends on an external knowledge base which contains new or correct knowledge. Aiming at amending the output of LLMs, a new knowledge base will be connected with the base model to implement a retrieve for needed new knowledge to a prompt or a question (Murty et al., 2022; Mitchell et al., 2022; Li et al., 2022; Madaan et al., 2022). Mitchell et al. (Mitchell et al., 2022) store manual edits in a memory module, and use a classifier to call the knowledge stored in the memory. Madaan et al. (Madaan et al., 2022) leverage the memory of user's feedback to generate prompts for LLMs. Instead of gradient calculation, Zheng et al. (Zheng et al., 2023) utilize the in-context learning method to revise the output of LLMs with demonstrations extracted from the corpus based on similarity.

Adding Additional Parameters refers to injecting a few trainable parameters which represent new knowledge to LLMs while original parameters keeping frozen (Dong et al., 2022; Huang et al., 2022; Raunak and Menezes, 2022; Dai et al., 2023). Dong et al. (Dong et al., 2022) put forward a lightweight feed-forward network to add new parameters adapted to specific factual contexts for knowledge generalization. Huang (Huang et al., 2022) et al. design an editor called Transformer-Patcher, which is capable of modifying the mistake of LLMs sequentially by adding and training a few neurons in transformer.

b. The method modifying model's parameters

Fine-tuning is a general technique since pre-training model has been widely adopted in NLP research, which always obtains promising results in downstream tasks. Meanwhile, fine-tuning is an intuitive and effective method to urge the model to learn new knowledge for model editing (Zhu et al., 2020; Zhang et al., 2022; Yao et al., 2023). Recently, there are a series of parameter-efficient fine-tuning methods, such as Prefix-Tuning (Li and Liang, 2021) and LoRA (Hu et al., 2021), making it more appreciate for knowledge editing based on fine-tuning. Zhang et al. (Zhang et al., 2022) operate incremental parameter updates of different amounts by calculating the importance of the weight matrix to improve the update efficiency and adaptability. Zhu et al. (Zhu et al., 2020) leverage a loss constraint attached to the base model to reduce the impact on irrelevant knowledge during the process of fine-tuning. Similarly, Lee et al. (Lee et al., 2022) also implement large-scale continual learning for knowledge updating with regularized fine-tuning.

Meta-learning is aimed at updating the knowledge in LLMs through varying their parameters with the prediction from a well-trained hypyernetwork (Sinitsin et al., 2019; Mitchell et al., 2021; De Cao et al., 2021). Mitchell et al. (Mitchell et al., 2021) propose an auxiliary network with gradient decom-