



Figure 2: Objectives of the knowledge updating in large language model.

and the parameters of the original model  $f_\theta$ , respectively. Similarly, we first fine-tune the model  $f_\theta$  on a dataset containing old knowledge, and then subtract the parameters  $\theta$  of the original model  $f_\theta$  from model’s parameters after fine-tuning to obtain the knowledge parameters  $\theta_\Delta^{old}$  indicating the old knowledge, as follows:

$$\theta_\Delta^{old} = \text{FT}\{\theta, K_{old}\} - \theta, \quad (3)$$

where  $K_{old}$  refers to a dataset composed of old knowledge which we desire to forget. Inspired by (Ilharco et al., 2022), we believe that subtracting the parameters  $\theta_\Delta^{old}$  from  $\theta$  can assist the model  $f_\theta$  to forget this part of old knowledge. So we define the process of forgetting old knowledge as follows:

$$\theta' = \theta - \lambda\theta_\Delta^{old}, \quad (4)$$

where  $\lambda$  is a hyper-parameter to control the rate of forgetting. Now we gain a new model  $f_{\theta'}$  with its parameters  $\theta'$ , which has forgotten the old knowledge compared to  $f_\theta$ . Note that this process of forgetting old knowledge only makes sense if the model  $f_\theta$  has already learned the old knowledge, otherwise, there is no need for forgetting and the forgetting operation may have a destructive effect on the normal knowledge of the model.

## 4.2 Learning new knowledge

With the model  $f_{\theta'}$  that has gone through the process of forgetting old knowledge, then we will inject the new knowledge to  $f_{\theta'}$  for knowledge updating by supervised fine-tuning. Similarly, we define the process of learning new knowledge as follows:

$$\theta^* = \text{FT}\{\theta', K_{new}\} \quad (5)$$

where  $\text{FT}$  is the operation of fine-tuning,  $\theta^*$  is the parameters of a new model  $f_{\theta^*}$  which has learned the new knowledge compared to  $f_\theta$  and  $K_{new}$  refers to a dataset composed of new knowledge which we need to learn and update for  $f_\theta$ .

## 5 Experiments

### 5.1 Datasets

In this work, we use ZsRE (Levy et al., 2017) and COUNTERFACT (Meng et al., 2022a), two widely used datasets, for our experiments. ZsRE is a Question Answering (QA) dataset that utilizes question rephrasings generated by back-translation as the equivalence neighborhood. COUNTERFACT is a more challenging dataset with counterfactual data. We follow the setting of (Yao et al., 2023) to take the eval and edit sets of which there are 19,085 and 10,000 pieces of data respectively. Moreover, we divide the datasets into two parts of old knowledge and new knowledge respectively to achieve two-stage knowledge update. The following is an example of old knowledge and new knowledge in zsRE, which represents the modification of knowledge from "Los Angeles" to "New Orleans". More details about the datasets and examples can be found in the appendix A.1.

#### The old knowledge:

{ "instruction": "What city did Marl Young live when he died?", "input": "", "output": "Los Angeles" }

#### The new knowledge: