

Forgetting before Learning: Utilizing Parametric Arithmetic for Knowledge Updating in Large Language Models

Shiwen Ni^{♣*}, Dingwei Chen^{♣*}, Chengming Li^{◇†}, Xiping Hu[◇], Ruifeng Xu[♡], Min Yang^{♣†}

[♣] Sun Yat-Sen University

[♣] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

[◇]Shenzhen MSU-BIT University [♡] Harbin Institute of Technology (Shenzhen)

{sw.ni, min.yang}@siat.ac.cn, chendw26@mail2.sysu.edu.cn

{licm, huxp}@smbu.edu.cn, xuruifeng@hit.edu.cn

Abstract

Recent advancements in Large Language Models (LLMs) have showcased their remarkable capabilities in text understanding and generation. However, even stronger LLMs are susceptible to acquiring erroneous or obsolete information from the training corpus. Direct secondary fine-tuning with data containing new knowledge may be ineffective in updating knowledge due to the conflict between old and new knowledge. In this paper, we propose a new paradigm for fine-tuning called **F-Learning** (Forgetting before Learning), which employs parametric arithmetic to facilitate the forgetting of old knowledge and learning of new knowledge. Experimental results on two publicly available datasets demonstrate that our proposed F-Learning can obviously improve the knowledge updating performance of both full fine-tuning and LoRA fine-tuning, simultaneously outperforming the existing baselines in most cases. Moreover, we have also discovered that forgetting old knowledge by subtracting the parameters of LoRA can yield a similar effect to subtracting the parameters of full fine-tuning, and occasionally even surpass it significantly.

1 Introduction

Large Language Models (LLMs) possess an extraordinary ability to understand and generate natural language (Brown et al., 2020; Raffel et al., 2020; Ouyang et al., 2022). Although LLMs are very capable of learning, they are not immune to the acquisition of incorrect knowledge in the corpus. Moreover, much of the knowledge in the real world is constantly updated, and some of the originally correct knowledge in LLMs can become outdated and invalid over time. For example, the question "Who is the President of the United States?" is answered "Donald Trump" in the year 2020, while the answer

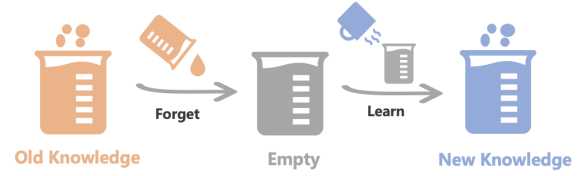


Figure 1: Diagram for "Forgetting before Learning".

now is "Joe Biden". Consequently, the challenge with LLMs is continuously updating to ensure they reflect current, correct knowledge. Existing methods of model editing and knowledge updating usually add additional network (Dong et al., 2022; Huang et al., 2022; Raunak and Menezes, 2022), model parameters (Dai et al., 2023; Dong et al., 2022; Huang et al., 2022), knowledge bases (Murty et al., 2022; Mitchell et al., 2022; Li et al., 2022; Madaan et al., 2022; Mitchell et al., 2022; Zheng et al., 2023), etc., and the editing process is not as straightforward and simple as fine-tuning methods (Zhang et al., 2022; Li and Liang, 2021; Hu et al., 2021) directly with new knowledge. Currently, the most used method for learning new knowledge is still direct fine-tuning of the model.

Empirically, when human beings establish their own initial cognition, if they are exposed to new knowledge that is inconsistent with their initial cognition, they usually feel conflicted and it is difficult for them to learn and accept the new knowledge. If the original cognition and knowledge are forgotten, then the new knowledge to be learned will not conflict with the original cognition and knowledge, which makes it better to learn and absorb the new knowledge. As shown in Figure 1, it is better to pour in the "new water" only after the "original water" in the cup has been poured out. For example, if people have been educated to believe that "the Earth is flat" since childhood, it would be challenging for them to accept the conflicting knowl-

*Equal contribution.

† Corresponding author.