the rate of forgetting increases. And under the same circumstances, LoRA brings less knowledge forgetting than the full fine-tuning.

6 Case Study

To further illustrate the effectiveness of the proposed method, we present a case study on the results of the knowledge updating by the original model, only forgetting old knowledge and performing F-learning. We selected some cases in the experiment of llama2-7B on zsRE dataset, noting that the hyper-parameters λ of the rate of forgetting is set to 0.3. Table 4 shows the results during different knowledge updating stages. From the first example and second example, we can find that model begins to output some irrelevant content after performing the forgetting operation, indicating that it gradually forgets the old knowledge. In example 4, the forgetting operation failed to assist the model in forgetting old knowledge, but it still completed knowledge updating with the help of F-learning. However, sometimes there are some bad cases, such as example 3, where the model never learned new knowledge, which shows that our method has certain limitations and could be improved.

7 Conclusion

In this paper, we propose a new paradigm of knowledge updating during supervised fine-tuning called **F-Learning** (Forgetting before Learning), which is based on parametric arithmetic to forget old knowledge and learn new knowledge for eliminating contradictions between old and new knowledge. The experiments on zsRE and COUNTERFACT datasets show that our method surpasses other baselines in most cases. Simultaneously we find that forgetting old knowledge by subtracting the parameters of LoRA can achieve the similar effect of subtracting the parameters of full fine-tuning, which is inspiring. We will further investigate the updating of knowledge.

8 Limitations

In this work, the proposed F-learning paradigm, although it improves the effectiveness of the fine-tuning methods for updating the knowledge of large language models, adds extra computation due to an extra forgetting process.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Damai Dai, Wenbin Jiang, Qingxiu Dong, Yajuan Lyu, and Zhifang Sui. 2023. Neural knowledge bank for pretrained transformers. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 772–783. Springer.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing commonsense knowledge in gpt. *arXiv preprint arXiv:2305.14956*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2022. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations*.