OVERCOMING CATASTROPHIC FORGETTING: A NOVEL FINE-TUNING METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite remarkable advances in Large Language Models (LLMs), a persistent challenge remains: the potential for these models to acquire erroneous or outdated information from their training data. Direct fine-tuning with data containing new knowledge can be ineffective due to conflicts between old and new knowledge. This paper proposes a novel fine-tuning paradigm called Delicate Fine-Tuning (**DFT**) that leverages parametric arithmetic to pinpoint the location of knowledge and update only the minimal set of relevant parameters. Experimental results on two publicly available datasets demonstrate that our proposed DFT significantly improves the knowledge updating performance of full fine-tuning, consistently outperforming existing baselines in most cases.

1 INTRODUCTION

Large Language Models (LLMs) exhibit remarkable proficiency in understanding and generating natural language Brown et al. (2020); Raffel et al. (2020); Ouyang et al. (2022). Despite their impressive learning capabilities, LLMs are susceptible to acquiring inaccurate knowledge from their training corpora. Furthermore, the dynamic nature of real-world knowledge necessitates continuous updates, as information that was once accurate may become outdated or invalid over time.

For instance, in 2020, the query ”Who is the President of the United States?” would have yielded ”Donald Trump” as the answer. However, the current answer is ”Joe Biden.” This exemplifies the ongoing challenge faced by LLMs: the need for continuous updating to ensure they reflect accurate and up-to-date knowledge.

Current approaches to model editing and knowledge updating typically involve augmenting the network architecture (Dong et al., 2022; Huang et al., 2022; Raunak & Menezes, 2022), introducing additional model parameters (Dai et al., 2023; Dong et al., 2022; Huang et al., 2022), or integrating external knowledge bases (Dai et al., 2023; Dong et al., 2022; Huang et al., 2022). These methods often necessitate more complex procedures than straightforward fine-tuning with new knowledge (Zhang et al., 2022; Li & Liang, 2021; Hu et al., 2021).

At present, direct fine-tuning of the model remains the predominant method for incorporating new knowledge.

During human cognitive development, individuals often encounter situations where new knowledge conflicts with their existing understanding.

They usually remember both the new knowledge and the old knowledge simultaneously, and then often get confused, leading to contradictions that make it difficult to learn the new knowledge. If we directly modify the memory of old knowledge and original cognition, 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. 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 knowledge that ”the Earth is round” when they become adults. Conversely, if they could directly modify their memory of the erroneous knowledge ”the Earth is flat” to the correct knowledge ”the Earth is round,” it would be much simpler.

So how do we locate the position of old knowledge and then update it accurately? Our research has shown that when fine-tuning large language models, they tend to learn sentence structure, grammar, and style first, with knowledge being acquired last. Therefore, we control the variables to prevent the model from learning sentence structure and stylistic information.

Inspired by the above empirical observations and (Ilharco et al., 2022)'s task arithmetic, we propose a novel paradigm of knowledge updating called **DFT** (Delicate Fine-Tuning). Specifically, DFT begins by using the large language model to predict and generate an answer, resulting in a data point. Next, DFT modifies only the key knowledge within the sentence, keeping the sentence structure and style intact, creating a new data point. We then fine-tune the model separately with both data points, recording the parameter changes. By comparing these parameter changes, we identify sections that exhibit similar changes in direction. These sections, representing aspects that are not relevant to the knowledge update, are discarded entirely.

We retain only the parameters exhibiting contrasting change directions, then compare their differences, rank those differences, and identify the top T % with the largest differences .Then update the top T % of parameters , where T is a predefined threshold ratio. The whole process is repeated iteratively until the model's knowledge update is complete.

This paper makes the following contributions:

- We propose a novel fine-tuning paradigm "**DFT** (Delicate Fine-Tuning)" for knowledge updating in large language models.
- Our experimental results show that **DFT** (Delicate Fine-Tuning) improves the knowledge updating performance across various fine-tuning methods and surpasses existing baselines in most cases.

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. (2023b):

a. The method preserving model's parameters

Adding Additional Parameters.

This approach involves injecting a small number of trainable parameters, representing new knowledge, into the LLM while keeping its original parameters frozen . This technique, explored by Dong et al. (2022); Huang et al. (2022); Raunak & Menezes (2022); Dai et al. (2023), allows for efficient knowledge injection without retraining the entire model. Dong et al. (2022) proposed a lightweight feed-forward network that incorporates additional parameters specifically tailored to factual contexts, enabling knowledge generalization.Huang et al. (2022) developed a model editor named Transformer-Patcher, which sequentially corrects errors in LLM outputs by adding and training a limited number of neurons within the transformer architecture.

Retrieve augmentation.

These methods rely on an external knowledge base containing new or corrected information, aiming to amend the output of LLMs by incorporating retrieved knowledge relevant to the given prompt or question. This approach, explored by Murty et al. (2022); Mitchell et al. (2022); Li et al. (2022); Madaan et al. (2022), facilitates the integration of new knowledge into the model's responses.

Mitchell et al. (2022) propose a memory module that stores manual edits, enabling a classifier to retrieve and apply the relevant knowledge.Madaan et al. (2022) leverage the memory of user feedback to generate prompts that guide LLMs toward more accurate responses. Alternatively, Zheng et al. (2023) utilize in-context learning to revise LLM outputs by extracting demonstrations from a corpus based on similarity, eliminating the need for gradient calculations.

b. The method modifying model's parameters

Fine-tuning has become a ubiquitous technique in NLP research, owing to the widespread adoption of pre-trained models for downstream tasks. Its intuitive nature and effectiveness in imparting new knowledge make it a valuable tool for model editing Zhu et al. (2020); Zhang et al. (2022); Yao et al.

(2023). Recent advancements in parameter-efficient fine-tuning methods, such as Prefix-Tuning Li & Liang (2021) and LoRA (Hu et al. (2021)), have further enhanced its applicability to knowledge editing. Zhang et al. (2022) proposed an adaptive fine-tuning strategy that dynamically adjusts the magnitude of parameter updates based on the importance of the weight matrix, thereby improving efficiency and adaptability. Zhu et al. (2020) introduced a loss constraint that minimizes the impact on irrelevant knowledge during fine-tuning, preserving the integrity of the base model. Similarly, Lee et al. (2022) explored large-scale continual learning for knowledge updating through regularized fine-tuning.

Meta-learning approaches aim to update knowledge within LLMs by adjusting their parameters based on predictions from a well-trained hypernetwork. This technique, investigated by Sinitsin et al. (2019); Mitchell et al. (2021); De Cao et al. (2021), enables efficient knowledge updates without retraining the entire model. Mitchell et al. (2021) introduced an auxiliary network with gradient decomposition, enabling efficient edits to LLMs based on a single input-output pair.De Cao et al. (2021) proposed updating specific weights within a subset of modules using a hypernetwork with constrained optimization.

Locate and edit targets the internal mechanisms of LLMs, aiming to modify specific parameters and neurons to correct outputs based on knowledge-driven interventions Meng et al. (2022a); Dai et al. (2022); Meng et al. (2022b); Santurkar et al. (2021); Geva et al. (2022). Geva et al. (2021) discovered that the feed-forward network layers within transformers store key-value pairs associated with specific knowledge. Meng et al. (2022a) employed a causal reasoning method to identify key neuron activations and update factual associations by modifying feed-forward weights. To facilitate large-scale knowledge editing, they introduced Meng et al. (2022b), a method that directly updates thousands of memories within LLMs. Gupta et al. (2023) enhanced knowledge updating by optimizing edit token selection and layer selection during the editing process. Yu et al. (2023) utilized partitioned gradients to identify significant weights for unlearning biases in the model.

Hiyouga hiyouga (2023) developed the fastedit software framework, which enables convenient editing of models using causal reasoning. Zhang et al. Zhang et al. (2024); Wang et al. (2023a); Yao et al. (2023); Cheng et al. (2023); Mao et al. (2023); Zhang et al. (2023) developed the EasyEdit software framework, which makes it easy to use a variety of methods for editing models.

While numerous methods have been proposed for knowledge updating in LLMs, many necessitate the introduction of additional knowledge bases, neural network modules, or model parameters. This often leads to practical challenges, including increased model complexity and inference costs. Therefore, this paper focuses on enhancing and refining fine-tuning methods as a more efficient and practical approach to knowledge updating in LLMs.

3 Task Definition

This paper addresses the task of knowledge updating in large language models (LLMs). Given a pre-trained model f_θ and a set of input-output knowledge pairs $K_{old} = (x_1, y_1), (x_2, y_2), ..., (x_i, y_i)$, the objective is to modify the model parameters θ to obtain a new model f_{θ^*} that generates a corresponding set of updated input-output pairs $K_{new} = (x_1, y_1^{new}), (x_2, y_2^{new}), ..., (x_i, y_i^{new})$. Here, i represents the number of knowledge pairs to be updated.

Following the definition in (Yao et al., 2023), we can formally express this process and its objective as:

$$f_{\theta^*}(x_i) = \begin{cases} y_i^{new} & \text{if } x_i \in N(x_i) \\ f_\theta(x_i) & \text{if } x_i \in other \end{cases} \tag{1}$$

where $N(x_i)$ represents x_i itself and its equivalent neighbourhood.

The knowledge update task aims to modify the model's responses only for x_i and its equivalent domain $N(x_i)$, where $N(x_i)$ represents the neighborhood of x_i encompassing semantically equivalent instances. The goal is to update the answers associated with x_i and its equivalent domain without affecting the responses to other out-of-scope knowledge.

The effectiveness of knowledge updating is evaluated based on the following three metrics:

a. Reliability

Measured as the average accuracy of the updated model f_{θ^*} on the new knowledge. This metric assesses the effectiveness of the update process itself. For example, the answer to the question ”Who is the President of the US?” should be updated from ”Donald Trump” to ”Joe Biden” after knowledge updating.

b. Generalization

Evaluated by the average accuracy of f_{θ^*} on examples drawn uniformly from the equivalence neighborhood $N(x_i)$. This metric assesses the ability of the model to generalize the update to semantically equivalent inputs. For example, the answer to the question ”Who holds the position of the President of the US?” should also be updated from ”Donald Trump” to ”Joe Biden”.

c. Locality

Assessed by the proportion of predictions from the updated model f_{θ^*} that remain unchanged compared to the pre-update model f_θ on irrelevant examples. This metric evaluates the ability of the model to preserve the original knowledge base while updating specific knowledge. For example, the answer to the question ”'You're fired!' is the catchphrase of which celebrity?” should remain unchanged as ”Donald Trump” after the update.

4 Proposed method: **DFT**

This section details our proposed approach for knowledge updating in LLMs. Departing from methods that rely on external knowledge bases or additional parameters, our method leverages a full fine-tuning strategy. The process is structured in two distinct stages:

4.1 Locate the parameters associated with the old knowledge

Supervised fine-tuning (SFT) on a designated dataset enables us to identify the direction of parameter alignment with the desired knowledge. This alignment is reflected in the variations observed in the model's parameters during the training process. Within this framework, we define incremental parameters, denoted as θ_Δ, as knowledge parameters for a given large language model f_θ and its parameters θ. These knowledge parameters are computed as follows:

$$\theta_\Delta = \mathrm{FT}\{\theta, \mathrm{K}\} - \theta \tag{2}$$

where FT is the operation of supervised fine-tuning, while K, θ refer to the dataset of knowledge and the parameters of the original model f_θ, respectively.

Analogously, we initially fine-tune the model f_θ on a dataset comprising the model's original knowledge. Subsequently, we subtract the original model parameters θ from the parameters obtained after fine-tuning to derive the knowledge parameters θ_Δ^{old}, representing the learned original knowledge. This calculation is expressed as:

$$\theta_\Delta^{old} = \mathrm{FT}\{\theta, \mathrm{K}_{\mathrm{old}}\} - \theta \tag{3}$$

where K_{old} refers to a dataset composed of the model's original knowledge. The related work in Ilharco et al. (2022) considers that subtracting the parameters θ_Δ^{old} from θ can assist the model f_θ to forget this part of old knowledge:

$$\theta' = \theta - \lambda \theta_\Delta^{old}, \tag{4}$$

where λ is a hyper-parameter to control the rate of forgetting. This process yields a new model, $f_{\theta'}$, with parameters θ', which exhibits reduced retention of the original knowledge compared to the initial model f_θ. The forgetting operation may have a destructive effect on the normal knowledge of the model.

However, we believe that $\theta_{\triangle}^{old}$ also contains other information such as sentence structure, grammar, and style, which requires further processing to accurately pinpoint the old knowledge.

Then, we re-fine-tune the model f_θ on a dataset containing new knowledge, and then subtract the parameters θ of the original model f_θ from model's parameters after fine-tuning to obtain the knowledge parameters $\theta_{\triangle}^{old}$ indicating the new knowledge, as follows:

$$\theta_{\triangle}^{new} = \mathrm{FT}\{\theta, \mathrm{K}_{\mathrm{new}}\} - \theta \tag{5}$$

where K_{new} refers to a dataset composed of new knowledge .

Then, We compare $\theta_{\triangle}^{old}$ and $\theta_{\triangle}^{new}$, discarding all elements with the same sign. Then, we select the top T % of elements in $\theta_{\triangle}^{new}$ with the largest difference from $\theta_{\triangle}^{old}$. T is a predefined threshold ratio.

$$\theta_{\triangle}^{core} = \mathrm{f}\{\theta_{\triangle}^{\mathrm{new}}, \theta_{\triangle}^{\mathrm{old}}, \mathrm{T}\} \tag{6}$$

$\theta_{\triangle}^{core}$ is the crucial parameter that needs to be updated.

4.2 Learning new knowledge by updating only the most relevant parameters

We define the process of learning new knowledge as follows:

$$\theta^* = \theta + \lambda\theta_{\triangle}^{core} \tag{7}$$

where λ is a hyper-parameter to control the rate of learning. We repeat the processes outlined in equations (3), (5), (6), and (7) until the model's output reflects the new knowledge. Now we gain a new model f_{θ^*} with its parameters θ^*, which has forgotten the old knowledge compared to f_θ. It learns only the new knowledge, avoiding any other information, preventing catastrophic forgetting caused by style changes and the like.

5 Experiments

5.1 Datasets

Our experiments employ two widely used datasets: Levy et al. (2017) and CounterFact (Meng et al., 2022a). ZsRE is a Question Answering (QA) dataset that leverages question rephrasings generated via back-translation to represent the equivalence neighborhood. CounterFact presents a more challenging benchmark with counterfactual data. Following the experimental setup outlined in Yao et al. (2023), we utilize the evaluation (eval) and edit sets of these datasets, comprising 19,085 and 10,000 data points, respectively. To facilitate two-stage knowledge update, we partition both datasets into sets of old knowledge and new knowledge. For instance, in ZsRE, a typical knowledge update scenario involves modifying the answer from ”Los Angeles” to ”New Orleans”, as illustrated in the following example:

The old knowledge:

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

The new knowledge:

{**”instruction”**: ”What city did Marl Young live when he died?”, **”input”**: ””, **”output”**: ”New Orleans” }

5.2 Baselines

To evaluate the effectiveness of the proposed DFT method, we conducted experiments comparing it to both fine-tuning methods and locate-based methods. For fine-tuning methods, we first compared DFT to full fine-tuning (Full-FT) and LoRA (Hu et al., 2021). LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning technique that introduces trainable low-rank matrices to the model's layers, enabling efficient adaptation while preserving the majority of the model's parameters. We

Table 1: Results on three metrics of the two datasets based on LLAMA2-7B and LLAMA-7B.

Dataset	**Editor**	**LLAMA2-7B**			**LLAMA-7B**		
		Reliability	**Generality**	**Locality**	**Reliability**	**Generality**	**Locality**
ZsRE	Original model	43.81	43.26	/	43.35	42.92	/
	LoRA	43.22	42.40	70.91	46.90	45.98	75.99
	DFT$_{\text{LoRA}}$	**48.73**	**48.14**	**76.23**	**49.95**	**49.22**	**78.54**
	FT-c	49.23	47.12	67.48	47.54	45.60	68.25
	Full-FT	81.23	74.96	70.62	70.83	66.80	65.37
	ROME	43.67	42.84	**93.85**	44.36	43.23	**99.51**
	MEMIT	83.78	79.35	70.61	78.21	77.74	69.23
	DFT$_{\text{FT}}$	**88.32**	**83.63**	74.44	**85.55**	**84.32**	**75.42**
CounterFact	Original model	18.68	17.06	/	21.72	17.96	/
	LoRA	30.67	23.33	40.19	27.63	21.32	39.86
	DFT$_{\text{LoRA}}$	**35.33**	**31.42**	**48.42**	**34.34**	**28.22**	**49.42**
	FT-c	29.51	19.77	19.52	26.72	17.88	20.10
	Full-FT	66.23	44.32	28.62	32.34	32.41	32.52
	ROME	18.61	17.43	**93.79**	21.99	19.32	**92.32**
	MEMIT	62.16	37.62	22.23	**57.12**	31.62	25.92
	DFT$_{\text{FT}}$	78.53	52.42	36.39	63.51	**45.44**	39.13

further investigated a fine-tuning approach with an L_∞ constraint (FT-c) (Zhu et al. (2020)), designed to retain irrelevant knowledge.

Regarding locate-based methods, we experimented with ROME (Meng et al. (2022a)), a method that updates specific factual associations through causal intervention. Finally, we compared DFT with MEMIT (Meng et al. (2022b)), a method known for its effectiveness in directly updating large-scale memories within LLMs.

5.3 Completion Details

For our experiments, we employ LLAMA2-7B and LLAMA-7B as the base models. The primary focus of our evaluation is the ability to update old knowledge with new knowledge. Therefore, we fine-tuned the base model using full fine-tuning for 3 epochs on the old knowledge dataset, resulting in the ”original model” for our experiments (akin to the ”original model” used in other studies). To maintain output consistency, we utilize the greedy decoding strategy during testing. Our experiments were conducted on a hardware platform comprising 8 x A800-80G GPUs.

5.4 Experimental Results

Table 1 presents the experimental results, our DFT method consistently outperforms other baselines in most cases. Notably, FT-c exhibits only minor improvements over the original model, potentially due to its reliance on norm regularization, which tends to preserve a portion of the old knowledge during the update process. As our original model has already acquired a substantial amount of old knowledge, learning new knowledge poses greater challenges.

Surprisingly, ROME maintains near-identical Reliability and Generalization scores compared to the original model on both datasets, while achieving high locality (exceeding 90%). This suggests limited knowledge updating by ROME, as the injection of new knowledge typically impacts locality. The limited parameter modification capability of ROME, combined with the extensive pre-existing knowledge in our original model, likely hinders the effectiveness of its causal tracing mechanism. It is noteworthy that full fine-tuning demonstrates a significantly greater capacity for learning new knowledge compared to LoRA. This difference can be attributed to LoRA’s focus on training a restricted subset of parameters within the attention structure, while a substantial portion of factual knowledge is encoded within the MLP layers.

Table 2: Results on three m55etrics of the zsRE dataset based on BLOOM-7B.

Editor	**Metric**		
	Reliability	**Generality**	**Locality**
Original model	28.02	27.95	/
LoRA	29.32	29.31	77.32
DFT$_{\text{LoRA}}$	**30.38**	**31.02**	**79.64**
Full-FT	44.32	43.72	63.94
DFT$_{\text{FT}}$	**45.83**	**44.64**	**72.15**

5.5 Updating with LoRA

Within this experimental framework, our approach involves simultaneous knowledge updating via full fine-tuning (or LoRA) in a single training process. We formally define this LoRA integrated approach as follows:

$$\theta_{\Delta}^{old} = \text{LoRA}\{\theta, \text{K}_{\text{old}}\} - \theta, \tag{8}$$

$$\theta_{\Delta}^{new} = \text{LoRA}\{\theta, \text{K}_{\text{new}}\} - \theta \tag{9}$$

$$\theta_{\Delta}^{core} = \text{f}\{\theta_{\Delta}^{\text{new}}, \theta_{\Delta}^{\text{old}}\} \tag{10}$$

$$\theta^{*} = \theta + \lambda\theta_{\Delta}^{core} \tag{11}$$

where LoRA represents the operation of supervised fine-tuning utilizing the LoRA technique . θ^* is noted as the parameters of the edited model f_{θ^*} which has completed the knowledge updating.

As presented in Table 1, the experimental results indicate that knowledge updating using LoRA outperforms full fine-tuning in certain instances. This improvement can be attributed to the parameter-efficient nature of LoRA-based knowledge forgetting, enabling more efficient learning and adaptation.

Empirical evidence from our experiments suggests that updating the model parameters through LoRA adaptation effectively approximates the performance achieved by full fine-tuning.

We hypothesize that this observation stems from the distributed nature of knowledge encoding across multiple model parameters. LoRA modifies the patterns and relationships associated with the old knowledge embedded within the attention structure, which represents an implicit knowledge representation.

5.6 Adaptability Testing

To further assess the adaptability of our proposed method, we conducted experiments on the zsRE dataset using BLOOM-7B as the base model. We maintained the same experimental settings as previously described. The results, presented in Table 2, demonstrate the continued effectiveness of DFT.

5.7 Time Testing

To evaluate the efficiency of our proposed DFT method, we compared the editing time of various knowledge updating and model editing methods for different edit sizes. Employing LLAMA2-7B as our base model, we present the results in Table 3.

Analysis of the results in Table 3 reveals that fine-tuning-based methods consistently exhibit significantly lower editing times compared to locate-based methods. This disparity can be attributed to the increased complexity and time requirements associated with locating specific neurons and parameters in locate-based methods. Furthermore, ROME's limitation to single-datapoint edits, in contrast to the batch editing capabilities of other methods, further diminishes its efficiency. Among fine-tuning-based methods, FT-c demonstrates faster optimization due to its norm constraint.

DFT method, while requiring multiple backward passes and comparisons as a multi-stage knowledge updating approach, necessitates updating only a limited set of parameters. Consequently, DFT

Table 3: Editing time for 1 edit, 10 edits, 100 edits of the two dataset based on LLAMA2-7B.Run ROME with FastEdit.Run MEMIT with EasyEdit

Editor	**1 edit**		**10 edits**		**100 edits**	
	zsRE	**COUNTERFACT**	**zsRE**	**COUNTERFACT**	**zsRE**	**COUNTERFACT**
FT-c	0.59(s)	0.55(s)	5.73(s)	5.57(s)	56.13(s)	55.12(s)
ROME	2.76(s)	2.46(s)	27.9(s)	24.32(s)	285.23(s)	242.21(s)
MEMIT	612(s)	606(s)	6231(s)	6193(s)	61831(s)	61631(s)
Full-FT	0.78(s)	0.74(s)	7.92(s)	7.43(s)	76.72(s)	75.11(s)
DFT$_{\text{FT}}$	**1.49(s)**	**1.42(s)**	**11.98(s)**	**11.21(s)**	**120.92(s)**	**118.87(s)**

exhibits an editing time approximately twice that of Full-FT, yet remains notably fast and convenient.

Further acceleration of supervised fine-tuning can be achieved through the utilization of deepspeed or other analogous optimization techniques.

5.8 Parametric Analysis of Updating Knowledge

DFT knowledge updating method hinges on the precise identification of knowledge-related parameters within the model. From an interpretability perspective, this approach allows us to pinpoint specific parameters containing the desired knowledge, enabling targeted updates. Furthermore, we conducted an in-depth analysis of the parameter distribution and its modifications within the LLMs.

Analysis reveals that parameter modifications in the MLP layers are more pronounced than those observed in the attention layers. This observation suggests that knowledge is primarily encoded within the MLP layers of the model.

6 Conclusion

This paper introduces a novel paradigm for knowledge updating during supervised fine-tuning, termed DFT (Differential Fine-Tuning). DFT leverages parametric arithmetic to pinpoint the location of existing knowledge and facilitates the acquisition of new knowledge, effectively resolving potential contradictions between old and new information.

Experimental evaluations conducted on the zsRE and CounterFact datasets demonstrate the superior performance of our proposed method compared to other baselines in most scenarios.

7 Limitations

While the proposed DFT paradigm enhances the efficacy of fine-tuning methods for updating knowledge in large language models, it incurs an increase in computational overhead due to the incorporation of multiple backward passes.

References

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

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*, 2023.

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

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

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

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

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

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 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*, pp. 30–45, 2022.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

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

hiyouga. Fastedit: Editing llms within 10 seconds. `https://github.com/hiyouga/FastEdit`, 2023.

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

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2022.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. Plug-and-play adaptation for continuously-updated qa. *arXiv preprint arXiv:2204.12785*, 2022.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*, 2017.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*, 2022.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2833–2861, 2022.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Editing personality for llms. *arXiv preprint arXiv:2310.02168*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2022b.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2021.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022.

Shikhar Murty, Christopher D Manning, Scott Lundberg, and Marco Tulio Ribeiro. Fixing model bugs with natural language patches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11600–11613, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Vikas Raunak and Arul Menezes. Rank-one editing of encoder-decoder models. *arXiv preprint arXiv:2211.13317*, 2022.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. *Advances in Neural Information Processing Systems*, 34:23359–23373, 2021.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *International Conference on Learning Representations*, 2019.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*, 2023a.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*, 2023b.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, 2023.

Ningyu Zhang, Jintian Zhang, Xiaohan Wang, Honghao Gui, Kangwei Liu, Yinuo Jiang, Xiang Chen, Shengyu Mao, Shuofei Qiao, Yuqi Zhu, Zhen Bi, Jing Chen, Xiaozhuan Liang, Yixin Ou, Runnan Fang, Zekun Xi, Xin Xu, Lei Li, Peng Wang, Mengru Wang, Yunzhi Yao, Bozhong Tian, Yin Fang, Guozhou Zheng, and Huajun Chen. Knowlm technical report, 2023. URL `http://knowlm.zjukg.cn/`.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2022.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*, 2023.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*, 2020.