Semantic are Beacons: A Semantic Perspective for Unveiling Parameter-Efficient Fine-Tuning in Knowledge Learning

Renzhi Wang[1,2], **Piji Li**[1,2*]
[1] College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, China
[2] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[1]{rzhwang,pjli}@nuaa.edu.cn

Abstract

Parameter-Efficient Fine-Tuning (PEFT) methods enable efficient adaptation of Large Language Models (LLMs) to various downstream applications. However, the effectiveness of the PEFT diminishes notably when downstream tasks require accurate learning of factual knowledge. In this paper, we adopt a semantic perspective to investigate this phenomenon, uncovering the reasons behind PEFT's limitations in knowledge learning task. Our findings reveal that: (1) PEFT presents a notable risk of pushing the model away from the intended knowledge target; (2) multiple knowledge interfere with each other, and such interference suppresses the learning and expression of knowledge features. Based on these insights, we introduce a data filtering strategy to exclude data that is detrimental to knowledge learning and a re-weighted learning strategy to make the model attentive to semantic distance during knowledge learning. Experimental results demonstrate the effectiveness of the proposed method on open-source large language model, further validate the semantic challenge in PEFT, thus paving the way for future research.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capacity for comprehending and generating human-like text (OpenAI, 2023; Touvron et al., 2023a,b; Jiang et al., 2023; Zhao et al., 2023). Customizing a general-purpose LLM for specific tasks commonly involves fine-tuning all model parameters (He et al., 2022; Wang et al., 2023a; Ovadia et al., 2023; Ni et al., 2023). With the increasing size of models, Parameter-Efficient Fine-tuning (PEFT) approaches like Adapter-tuning (Houlsby et al., 2019) and LoRA (Hu et al., 2022) have been proposed due to constraints imposed by computational resources and the quantity of data available for downstream tasks (Lialin et al., 2023; Pu et al., 2023; Xu et al., 2023). These methods selectively fine-tune a limited number of (additional) model parameters while keeping the majority of original parameters fixed (Lialin et al., 2023; Pu et al., 2023). However, the effectiveness of PEFT diminishes notably when downstream tasks necessitate precise learning of factual knowledge, such as proper nouns, temporal information, or geographic locations (Allen-Zhu and Li, 2023; Ling et al., 2023).

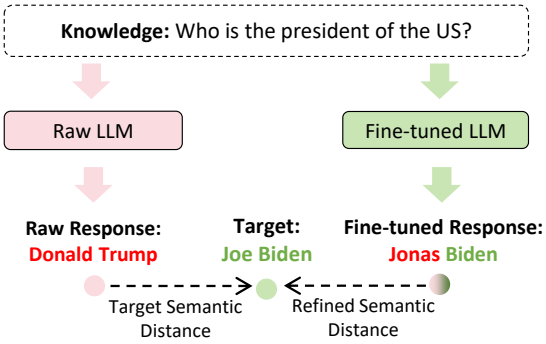

Figure 1: Illustration of semantic distance.

In this paper, we adopt a semantic perspective to examine the process of PEFT guiding model learning to elucidate potential factors leading to its suboptimal knowledge learning outcomes. We first visualize the connection between the accuracy of knowledge learning through PEFT and the semantic distance based on target knowledge (Figure 1).We observe that optimal learning outcomes occur within a certain appropriate range of semantic distances from the target knowledge, whereas significant declines in learning efficacy for both proximal and distal semantic distances.

To investigate the underlying causes of this phenomenon, we conducted experimental analyses from the perspectives of learning one piece of knowledge and learning multiple knowledge, which involves using one data point or multiple data points for training the model. For learning single knowledge, typically, model parameters are

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 9523–9537

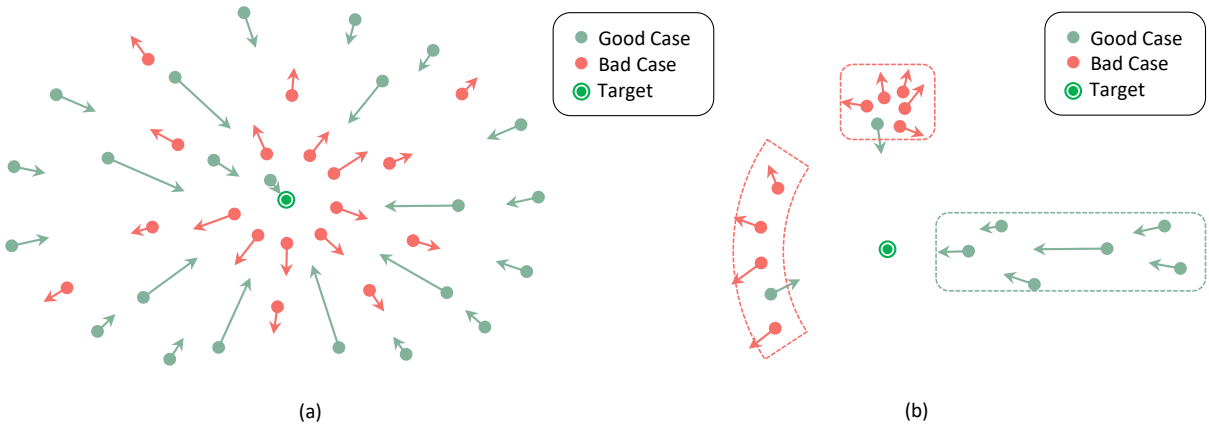

Figure 2: Illustration of (a) the away from target phenomenon in single knowledge and (b) knowledge interference across multiple knowledge in semantic space. Each dot represents original knowledge, with arrows indicating the knowledge learned by the model after training. The central target represents the normalized target knowledge. Green denotes the normal progression towards the target knowledge, red denotes an abnormal deviation away from the target. Dashed lines in (b) enclose multiple knowledge learned simultaneously.

iteratively updated during training via loss functions and gradient descent to gradually approach the target (Lv et al., 2017; Haji and Abdulazeez, 2021). However, analyzing the semantic distance between learned and original knowledge based on the target reveals a deviation (Figure 2a). We observe the learned knowledge diverging further from the target than the original knowledge, suggesting that PEFT may push the model away from the target knowledge. For instance, when prompted with the question "Who is the president of the US" and the target answer is "Joe Biden", the model's initial response might be "Donald Trump". While ideally, even partial learning of correct knowledge such as the last name "Biden" is acceptable, the model might instead produce responses with significantly greater semantic distance, such as "Jack Taylor" [1].

For multiple knowledge (Figure 2b), we analyze the relationship between fine-tuned parameters and original model parameters using Frobenius norm (Hu et al., 2022). Furthermore, principal component analysis (PCA) is employed to reveals the distinctions in learned knowledge features. Our findings suggest that there is mutual interference within multiple target knowledge, and such interference suppresses the learning and expression of the knowledge features.

Drawing on insights from the semantic perspective, we explore two concise methods to enhance PEFT's efficiency and effectiveness in knowledge learning: (1) A data filtering strategy, aimed at eliminating portions of the dataset that adversely affect overall knowledge learning. (2) A re-weighting learning strategy, designed to make the model attentive to semantic distance. Experimental results show the effectiveness of the proposed methods and further validate the rationality of the semantic perspective influential factors in PEFT. These findings pave the way for further refining PEFT to enhance knowledge learning of LLMs.

2 Semantic Perspective Analysis

This section investigates the factors influencing the effectiveness of model fine-tuning for knowledge learning from a semantic perspective. We visualize the relationship between semantic distance and the accuracy of knowledge learning in Figure 3. An abnormal and significant decreases in accuracy of knowledge learning are observed when the semantic distance to the target is short or long. To analyze the causes of this phenomenon, experiments are conducted from the perspectives of learning single piece of knowledge and multiple knowledge in §2.1 and §2.3, respectively.

2.1 Knowledge Learning and Semantic Distance

The task of large language model knowledge learning aims to adjust an initial model's behavior on the target knowledge (Mitchell et al., 2022b; Yao et al., 2023). Specifically, the raw model f_θ is represented by a function $f : \mathbb{X} \mapsto \mathbb{Y}$ that associates an input knowledge x with its corresponding

[1] Note that this is a demonstration example, where in some models the semantic distance between "Jack Taylor" and "Joe Biden" may be smaller compared to the semantic distance between "Donald Trump" and "Joe Biden".

prediction y (Yao et al., 2023). Given a set containing n items of input-output target knowledge pairs $K_{target} = \{(x_1, y_1^t), (x_2, y_2^t), ..., (x_n, y_n^t)\}$, the parameters of the model need to be updated to obtain a new model $f_{\theta^*}(x)$. Based on this, the internal knowledge of the raw model f_θ can be represented as $K_{old} = \{(x_1, y_1^{old}), (x_2, y_2^{old}), ..., (x_n, y_n^{old})\}$, where $y_i^{old} = \text{argmax}_y f_\theta(y \mid x_i)$. After tuning, the knowledge learned by the updated model $f_{\theta^*}(x)$ can be represented as $K_{new} = \{(x_1, y_1^{new}), (x_2, y_2^{new}), ..., (x_n, y_n^{new})\}$, similarly $y_i^{new} = \text{argmax}_y f_{\theta^*}(y \mid x_i)$.

From a semantic perspective, during training process, the model should gradually approach the target knowledge based on the loss function and gradient update strategy (Haji and Abdulazeez, 2021; Lv et al., 2017). When the model accurately learns the target knowledge ($y_i^{new} = y_i^t$), the semantic distance between the learned knowledge and target knowledge diminishes to 0. To distinguish different knowledge and measure the extent to which the model moves towards the target knowledge, we define the semantic distance between the original parameter knowledge (x_i, y_i^{old}) and the target knowledge (x_i, y_i^t) based on cosine similarity (referred to as "**Target Semantic Distance**"):

$$dist(y_i^{old}, y_i^t) = 1 - sim(\text{EMB}(y_i^{old}), \text{EMB}(y_i^t)) \quad (1)$$

where $\text{EMB}(y_i)$ represents the word embedding of y_i based solely on the embedding layer of the experimental LLM, and $sim(\cdot)$ represents cosine similarity. For handling multiple tokens, we opt for the mean pooling of these embedding. The semantic distance between the learned knowledge of the fine-tuned model (x_i, y_i^{new}) and the target knowledge (x_i, y_i^t) is similar to Formula 1. This semantic distance helps in analyzing the learning process of models when faced with different types of knowledge. A farther target semantic distance indicates a larger disparity between the knowledge to be learned and the original knowledge.

Following previous works (Huang et al., 2023a; Meng et al., 2022; Mitchell et al., 2022a; Cao et al., 2021), we employ the simple and straightforward evaluation metric of accuracy to explicitly measure the precision of knowledge learning. A higher accuracy indicates a more precise learning of target knowledge. The definition is as follows:

$$\mathbb{E}_{(x_i, y_i^t) \sim K_{target}} \mathbb{1}\left\{y_i^{new} = y_i^t\right\} \quad (2)$$

Experimental Settings We choose LLaMA2-7B (Touvron et al., 2023b), specifically its chat version, as our primary model for investigation, due to its popularity as an open-source model and its moderate model size that is suitable for our hardware resource. Additionally, considering models with varying structures and parameters, we also conducted experiments on BLOOMZ-1.7B/3B/7.1B (Muennighoff et al., 2023) and Mistral-7B (Jiang et al., 2023). For PEFT methods, we selected commonly used approaches in the era of LLMs such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), AdaLoRA (Zhang et al., 2023), and Adapter-tuning (Houlsby et al., 2019) to fine-tune the model, and compared them with the full fine-tuning.

For dataset, we use two prominent knowledge datasets: ZsRE (Levy et al., 2017) and COUNTERFACT (Meng et al., 2022), with their details available in Appendix B. In this experiment, 100 pieces of knowledge are sampled from the dataset for fine-tuning and evaluation each time, considering the limited practicality of learning a small number of knowledge (like 1 or 10) in real-world scenarios (we still conducted experiments with this volume of data to demonstrate the universality of the issues raised in Appendix D) and the low accuracy of learning more knowledge (Zhang et al., 2024). Moreover, our preliminary experiments show notable differences in the learning effectiveness of different knowledge samples (consistent with the analysis in §2.3). Therefore, we sample 100 times for each experimental to avoid incidental circumstances and obtain general patterns. Implementation details can be found in Appendix C.

Results and Analysis Figure 3 illustrates the learning performance of the model in the face of different target semantic distance knowledge. Overall, the larger target semantic distance, the lower knowledge learning accuracy, possibly indicating that larger semantic distances imply more challenging learning tasks. Compared to full fine-tuning, PEFT's knowledge learning performance is inferior, and there is a significant anomalous decrease in learning performance in situations where semantic distances are short or long regardless of the model type, the parameter size, and the amount of knowledge to be learned (Appendix D). We conducted experiments separately on single and multiple knowledge to analyze potential factors contributing to this phenomenon.

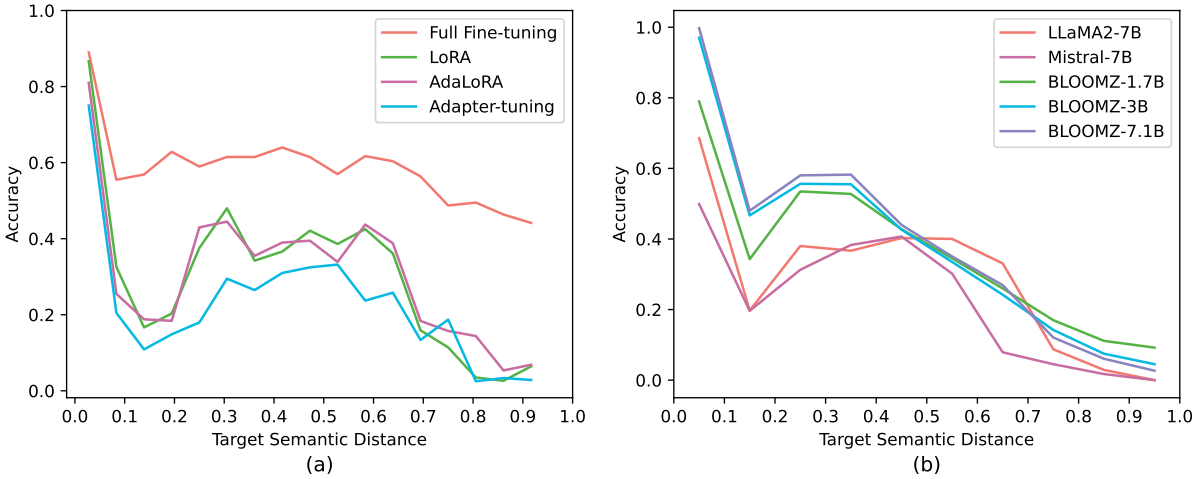

Figure 3: The relationship between knowledge learning accuracy and target semantic distance (a) based on model: LLaMA2-7B-chat, (b) based on method: LoRA. Target semantic distance is defined in Equation 1. Performance degradation occurs under short or long target semantic distance regardless of the method and model.

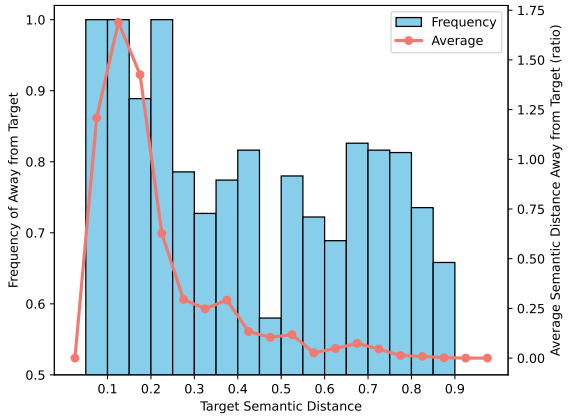

Figure 4: Relationship between deviation phenomenon and semantic distance. Note, on the far left and far right, there is blank space because the proportion there is 0.

Model	Proportion↑	
	LoRA	AdaLoRA
LLaMA2-7B	71.10	68.54
Mistral-7B	58.28	55.36
BLOOMZ-1.7B	63.03	59.17
BLOOMZ-3B	72.69	69.02
BLOOMZ-7B	81.63	79.49

Table 1: Proportion of deviation phenomenon in bad cases. Note, bad case is defined as accuracy score < 1 or $dist(y_i^{new}, y_i^t) > 0$.

2.2 Single Knowledge

In this section, we analyze the above phenomenon from the perspective of learning single knowledge. In §2.1, we expect the model to progress towards target knowledge during training process, guided by the loss function and gradient updates. Thus, when the model accurately acquires the target knowledge ($y_i^{new} = y_i^t$), the semantic distance between the learned and target knowledge becomes 0. However, the following experiments demonstrate that PEFT presents a significant risk of diverting the model away from intended knowledge target.

Experiments To demonstrate that PEFT pushes the model away from the target, we compute the target semantic distance between the newly learned knowledge y_i^{new} and the target knowledge y_i^t based on Equation 1. If the post-tuned semantic distance surpasses the target semantic distance ($dist(y_i^{new}, y_i^t) > dist(y_i^{old}, y_i^t)$), the model anomalously moves away from target knowledge.

We use the proportion of this phenomenon within the entire test cases to reflect the frequency of its occurrence, and the average deviation from the target to reflect the extent of its impact. Considering the initial differences in target semantic distance, we calculate the average relative deviation distance. For target knowledge y_i^t, old knowledge y_i^{old}, and newly acquired knowledge y_i^{new}, the definition of relative deviation distance from the target semantic distance is as follows:

$$RD = \frac{dist(y_i^{new}, y_i^t) - dist(y_i^{old}, y_i^t)}{dist(y_i^{old}, y_i^t)} \quad (3)$$

Based on this, we draw the relationship between this deviation phenomenon and target semantic distance shown in Figure 4. Additionally, by calculating the proportion of this anomaly occurring in inadequately learned knowledge (bade case, defined as accuracy score < 1 or $dist(y_i^{new}, y_i^t) > 0$) of dif-

ferent models, we further explore the universality of this phenomenon in Table 1.

To eliminate the mutual interference between knowledge, we sample one knowledge each time to fine-tune the model, with a total of 1000 samplings. Figure 4 is based on LLaMA2-7B-chat, LoRA and dataset ZsRE (also used in Table 1). Other configurations and details remain consistent with §2.1.

Results and Analysis Figure 4 illustrates the relationship between the deviation phenomenon and target semantic distance. On the one hand, this phenomenon consistently accounts for more than 50% across all test cases, indicating that the model not only has a high probability of not moving towards the target during training but may even be moving in the opposite direction. Moreover, this phenomenon occurs more frequently when the target semantic distance is both short and long (the probability approaching nearly 100% when the semantic distance is between 0.10~0.20), while the frequency decreases at moderate semantic distances.

On the other hand, the extent of this phenomenon decreases as the distance from the target increases, with instances where models deviate greatly even surpassing twice the semantic distance from the target in proximal cases. By comparing Figure 3 and Figure 4, it can be observed that the highest frequency and the deepest degree of deviation (semantic distance between 0.10~0.20) coincide with the significant decline in model knowledge learning performance, indicating a close association between this abnormal phenomenon and the decrease of model performance. Furthermore, Table 1 reveals that the proportion of this phenomenon consistently exceeds 50% regardless of the model and method, demonstrating its universality.

Regarding the possible reasons for this phenomenon, considering the success of countless neural networks training under the guidance of the loss functions and gradient descent strategies in the past, we believe that the model can be effectively directed towards the desired objectives based on these principles. So, the phenomenon of deviating from the target knowledge may be due to differences between the target of current loss function (cross-entropy loss) and the target of closer semantic distance we desire. Particularly when the target semantic distance falls within a relatively close range, this difference may become more pronounced, although both loss function and semantic distance become zero when the model accurately learns the knowledge. Based on this, we designed a simple re-weighting strategy to adjust the loss function during model training (§3.2), resulting in an improvement in the accuracy of knowledge learning, thus confirming this possible reason.

2.3 Multiple Knowledge

We proceed to investigate the phenomenon in §2.1 from the perspective of learning multiple knowledge. First, we analyze the relationship between the mean, variance of multiple knowledge semantic distances and learning accuracy. Then, we employ the Frobenius norm to analyze the relationship between fine-tuned parameters and original model parameters. Building upon this, we use PCA to reveal the distinctions in learned knowledge features, thus illustrating the mutual influences among multiple pieces of knowledge.

Experiments We utilize the mean semantic distance of multiple pieces of knowledge to characterize the average deviation of the learned knowledge from the target knowledge. Meanwhile, variance is employed to measure the dispersion of this phenomenon. By leveraging these two features, we aim to elucidate the relationship between the accuracy of the model in learning multiple knowledge and their target semantic distances.

We further delve into the relationship between fine-tuned parameters $\Delta\mathbf{W}$ and original parameters $\mathbf{W}$. Specifically, we inquire whether $\Delta\mathbf{W}$ exhibits a high correlation with $\mathbf{W}$, or mathematically, if $\Delta\mathbf{W}$ is predominantly aligned with the principal directions of $\mathbf{W}$. Additionally, we investigate the magnitude of $\Delta\mathbf{W}$ concerning its corresponding directions in $\mathbf{W}$.

Inspired by (Hu et al., 2022), we project $\mathbf{W}$ onto the r-dimensional subspace of $\Delta\mathbf{W}$ by computing $\mathbf{U}^\top\mathbf{W}\mathbf{V}^\top$, where $\mathbf{U}/\mathbf{V}$ represents the left/right singular-vector matrix of $\Delta\mathbf{W}$. Subsequently, we compare the Frobenius norm between $\|\mathbf{U}^\top\mathbf{W}\mathbf{V}^\top\|_F$ and $\|\mathbf{W}\|_F$. For comparison, we also calculate $\|\mathbf{U}^\top\mathbf{W}\mathbf{V}^\top\|_F$ by substituting $\mathbf{U}$ and $\mathbf{V}$ with the top k singular vectors of $\mathbf{W}$ or a random matrix. Considering the impact of semantic features on this relationship, we categorize models fine-tuned with different knowledge into three equidistant segments groups based on the mean semantic value of the target: close-distance learning, moderate-distance learning, and long-distance learning. We report the average norm of the query projection matrices in the self-attention module

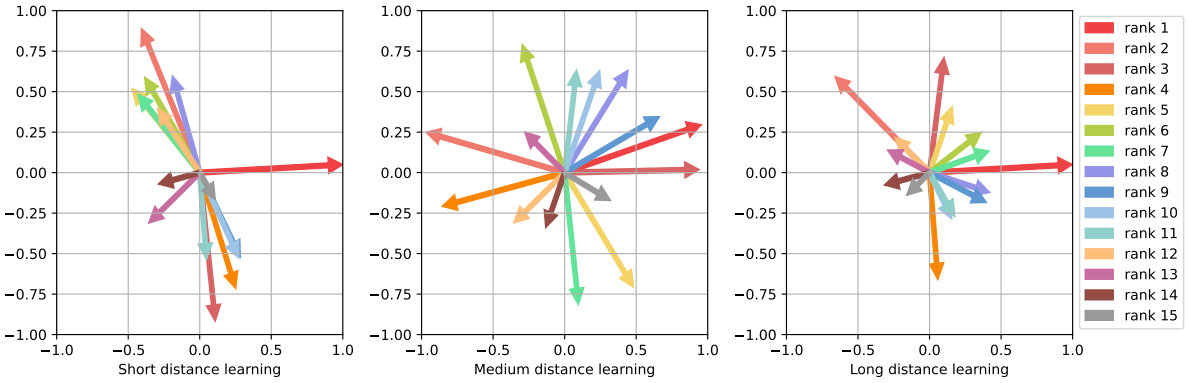

Figure 5: The PCA projection results of knowledge features at different semantic distances.

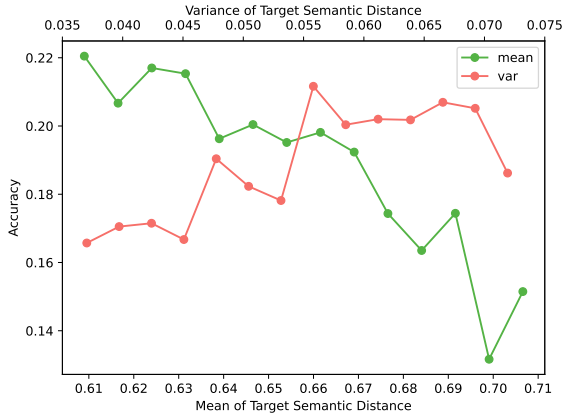

Figure 6: Relationship between mean/variance semantic distance and learning accuracy.

Norm	Short	Medium	Long
$\|\|\mathbf{W}_q\|\|_F$	88.4206	88.4206	88.4206
$\|\|\Delta\mathbf{W}_q\|\|_F$	0.1424	0.1564	0.1807
$\|\|\mathbf{U}_1^\top\mathbf{W}_q\mathbf{W}_1^\top\|\|_F(\Delta\mathbf{W}_q)$	0.1802	0.1638	0.2873
$\|\|\mathbf{U}_2^\top\mathbf{W}_q\mathbf{W}_2^\top\|\|_F(\mathbf{W}_q)$	1.3884	1.3884	1.3884
$\|\|\mathbf{U}_3^\top\mathbf{W}_q\mathbf{W}_3^\top\|\|_F(Random)$	0.0809	0.0832	0.08212

Table 2: The Frobenius norm of $\mathbf{U}_i^\top\mathbf{W}_q\mathbf{V}_i^\top$ where $\mathbf{U}_i$ and $\mathbf{V}_i$ are the left/right top r singular vector directions of either (1) $\Delta\mathbf{W}_q$, (2) $\mathbf{W}_q$, or (3) a random matrix.

($\mathbf{W}_q$) for each group in Table 2, considering $\Delta\mathbf{W}_q$ appears to have a higher intrinsic rank than the value projection matrices $\Delta\mathbf{W}_v$ (Hu et al., 2022).

The Frobenius norm provides a macroscopic analysis of the characteristic variations in model parameters. In order to conduct a more detailed comparison of the distinctions between multiple knowledge features mastered by different models, we employ PCA to compare the knowledge features in the fine-tuned models. We compute the first 15 principal components of the final layer outputs when the models are presented with various test knowledge inputs.

The configurations and details remain the same as §2.1. The experiments are based on the LoRA and ZsRE like §2.2. The weight matrices are all taken from the 32th layer of LLaMA2-7B-chat.

Results and Analysis In Figure 6, the accuracy of learning increases as the average semantic distance decreases and as the variance increases. This implies that there is mutual interference among different pieces of knowledge. And, the greater the distance between knowledge, the smaller the interference, thus the better learning accuracy outcome.

Table 2 shows that $||\mathbf{U}_1^\top\mathbf{W}_q\mathbf{W}_1^\top||_F(\Delta\mathbf{W}_q) > ||\mathbf{U}_3^\top\mathbf{W}_q\mathbf{W}_3^\top||_F(Random)$, indicating a stronger correlation of $\Delta\mathbf{W}_q$ with $\mathbf{W}_q$ compared to the random matrix. Leveraging properties of norms and the physical interpretation of projecting $\mathbf{W}_q$ onto the r-dimensional subspace of $\Delta\mathbf{W}_q$, higher values denote greater alignment of characteristic directions represented by the matrices, suggesting that $\Delta\mathbf{W}_q$ amplifies certain features already present in in $\mathbf{W}_q$. On the other hand, $||\mathbf{U}_1^\top\mathbf{W}_q\mathbf{W}_1^\top||_F(\Delta\mathbf{W}_q) < ||\mathbf{U}_2^\top\mathbf{W}_q\mathbf{W}_2^\top||_F(\mathbf{W}_q)$, with the former approximately 0.18 and the latter approximately 1.39, a difference of approximately 7.5 times. This suggests that the vector magnitude of $\mathbf{W}_q$ projected onto the r-dimensional subspace of $\Delta\mathbf{W}_q$ is smaller compared to $\mathbf{W}_q$ itself, implying a certain discrepancy in the amplification direction of $\Delta\mathbf{W}_q$ with respect to $\mathbf{W}_q$. This means, instead of merely reinforcing the top singular directions of $\mathbf{W}_q$, $\Delta\mathbf{W}_q$ can also amplify directions not emphasized in $\mathbf{W}_q$, indicating parameter-efficient fine-tuning can grasp certain knowledge that was absent in the original model. These findings consistent with those mentioned in (Hu et al., 2022). Futhermore, con-

cerning knowledge learning tasks, the amplification factor $\frac{\|\Delta \mathbf{W}\|_F}{\|\mathbf{U}^\top \mathbf{W} \mathbf{V}^\top\|_F}$ is not large. For knowledge of different semantic distances, the amplification factors are: 0.79 (short), 0.95 (medium), 0.63 (long). The amplification factor is higher in medium-distance learning condition and decrease in short/long scenarios, consistent with the performance variations in §2.1. In Figure 5, the knowledge features are richest in medium semantic distance learning, while the model's feature directions are restricted in short distances, and the features are too small in long distances (corresponding to the scenario where the amplification factor is minimal). This indicates that interference between knowledge can inhibit the direction and magnitude of knowledge features, thereby affecting the expression of knowledge, ultimately resulting in a decline in learning accuracy.

3 Applications of Our Semantic-Based Understanding

With insights from the semantic perspective, we propose two simple strategies from both data and model aspects to enhance the ability of PEFT in learning knowledge. We introduce a data filtering strategy in §3.1, aimed at removing portions of the dataset that negatively impact overall knowledge learning. In §3.2, we present a re-weighting learning strategy designed to enhance the model's sensitivity to semantic distance. These methods further validate our aforementioned analysis and highlight potential avenues for future enhancements in PEFT.

3.1 Data Filtering

From a data perspective, based on our analysis in §2, the accuracy of knowledge learning using currently PEFT methods is very low. This indicates that most knowledge has not been accurately learned, resulting in low data utilization. Moreover, there is mutual interference among the data, suggesting that we can discard or replace some "bad" knowledge. By reducing the interference between these knowledge, we can improve the accuracy of knowledge learning to some extent.

Method §2.3 points out that there is significant mutual interference among the data, and this interference hinders the learning and expression of knowledge features. Moreover, the accuracy of knowledge learning by the model increases as the semantic distance average decreases and as the variance increases. This suggests that, while keeping the test knowledge unchanged, we can improve learning effectiveness by selectively filtering and replacing training data to minimizing the average semantic distance while maximizing the variance, thereby reducing interference between knowledge and enhancing learning outcomes.

Specifically, given a knowledge dataset S, the initial target knowledge set for model fine-tuning $K_t = (x_i, y_i^t)$, we compute the mean $\mu(K_t)$ and variance $\sigma(K_t)$ based on target semantic distance defined by Formula 1. By continuously updating and optimizing the knowledge within K_t, we obtain the final training data knowledge set K_t^*:

$$\begin{aligned} &\arg\min_{K_t} \ \mu(K_t) - \lambda\, \sigma(K_t) \\ &s.t. \quad Mean_{\text{min}} < \mu(K_t) < Mean_{\text{max}} \end{aligned} \tag{4}$$

where λ is a non-negative weighting coefficient used to balance the relationship between variance and mean; $Mean_{\text{min}}$ and $Mean_{\text{max}}$ are given mean restriction ranges, considering that performance significantly decreases when the average semantic distance is too small. During the replacing process, at each step, we greedily remove the knowledge (x_k, y_i^k) from the current K_t to maximize the gain in the Formula 4, while simultaneously selecting knowledge $(x_k', y_i^{k'})$ from the dataset S which also maximizes the gain to replace it, until the desired proportion of replacement data is reached.

Experiments The experimental setup remains consistent with §2.1, where the model learns 100 items of knowledge sampled in the dataset each iteration, repeated 100 times. Only the training data is modified during the experiments, while the test data remain unchanged. We report the average accuracy of all the models. We conduct experiments on two datasets, ZsRE and COUNTERFACT, utilizing LoRA and LLaMA2-7B-chat. Using random data sampling from the dataset for replacement as a comparative experiment.

Results As shown in Figure 7, the proposed dataset filtering strategy effectively enhances the accuracy of model learning. Across various proportion of data update, the model's learning accuracy significantly outperforms random selection methods. Particularly, the best performance is achieved when 60% of the data is replaced. As the replaced data volume further increases, the improvement brought by data filtering begins to plateau, the improvement brought by data filtering starts to diminish, possibly due to the limited overlap between

Model	ZsRE (LoRA)			CounterFact (AdaLoRA)		
	Accuracy↑	Generality↑	Locality↑	Accuracy↑	Generality↑	Locality↑
BLOOMZ-1.7B	24.45 (+5.70)	22.50 (+3.25)	40.36 (-1.52)	25.53 (+4.55)	23.21 (+2.16)	40.42 (+1.29)
BLOOMZ-3B	20.49(+4.89)	17.96(+1.44)	42.20(+0.05)	22.37(+3.89)	18.68(+1.34)	41.73(-1.74)
BLOOMZ-7B	16.94(+3.13)	15.99(+1.42)	45.59(-0.34)	15.31(+2.86)	13.11(-1.20)	43.07(+2.30)
Mistral-7B	27.30(+5.07)	24.82(+1.88)	54.47(+2.09)	24.79(+4.93)	21.03(-1.24)	40.67(+3.40)
LLaMA2-7B	24.83(+5.29)	23.15(+1.91)	52.01(+0.15)	22.70(+5.91)	22.13(+2.24)	39.85(+1.74)

Table 3: The result of the re-weighting strategy. The values in parentheses represent the relative improvement/decrease compared to the original condition (the original results of the model are presented in Appendix F). We follow the same experimental settings as §2.2.

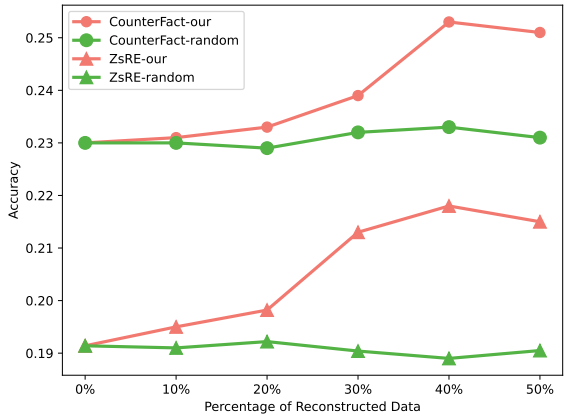

Figure 7: Effectiveness of data filtering strategy.

the training and testing knowledge caused by the excessive replacement of data.

3.2 Re-weighting Learning

We further explore a re-weighting learning strategy that allowing PEFT methods to aware the target semantic distance during training, thereby enhancing the effectiveness of knowledge learning.

Method In §2.2, we observed a significant likelihood for parameter-efficient fine-tuning to lead the model away from the target knowledge. Simultaneously, this anomaly is closely associated with a decrease in the model's learning accuracy. This deviation from the target knowledge implies that even with more training epochs, the model cannot accurately learn knowledge. We propose modifying the loss function to make the model aware of the target semantic distance, thus progressing along the correct fine-tuning direction during fine-tuning. By guiding the model towards the correct target knowledge, we aim to enhance the accuracy of knowledge acquisition.

Specifically, given the target knowledge (x_i, y_i^t), the original loss function computed based on the target knowledge during the training process (e.g., cross-entropy loss) is represented as $\mathcal{L}(y_i^t)$ and $\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}(y_i^t)$. We expect the model to pay more attention to both short and long semantic distance knowledge, thereby increasing the likelihood of learning correctly. Considering the target semantic distance defined in Formula 1 ranges from (0,1), the adjusted loss function based on the target semantic distance is as follows:

$$\mathcal{L}' = \mathcal{L} + \gamma \sum_{i=1}^{n} \lambda(y_i^t, y_i^{old}) \mathcal{L}(y_i^t) \quad (5)$$

$$\lambda(y_i^t, y_i^{old}) = 1 - \cos[dist(y_i^t, y_i^{old})\pi - \frac{\pi}{2}] \quad (6)$$

where γ serves as a non-negative weighting factor that adjusts the contribution of semantic loss to the overall loss function, and (x_i, y_i^{old}) is the original parameter knowledge. When the semantic distance of the target is moderate (close to 0.5), L' degenerates into the original loss function L (Wang et al., 2023b), resulting in minimal impact on regions where the original learning accuracy is relatively high.

Metrics In this section, we aim to assess the model's learning of knowledge from a broader perspective. Following (Yao et al., 2023), we employ the concepts of generality and locality to further evaluate the performance of knowledge learning. Generality refers to the expectation that a fine-tuned model f_{θ^*} should also adjust neighboring equivalents $N(x_i, y_i^t)$, such as rephrased sentences. The definition of generality is as follows:

$$\mathbb{E}_{(x_i', y_i') \sim N(x_i, y_i)} \mathbb{1} \left\{ \text{argmax}_y f_{\theta^*} \left(y \mid x_i' \right) = y_i' \right\} \quad (7)$$

Locality dictates that updates should be applied locally, ensuring that the fine-tuned model does not alter the outputs of irrelevant examples $O(x_i, y_i)$.

The definition of locality is as follows:

$$\mathbb{E}_{(x_i',y_i')\sim O(x_i,y_i)} \mathbb{1}\left\{f_{\theta^*}\left(y \mid x_i'\right) = f_\theta\left(y \mid x_i'\right)\right\} \tag{8}$$

where $N(x_i, y_i^t)$ is called neighboring equivalents, $O(x_i, y_i)$ is irrelevant examples, f_{θ^*} is the fine-tuned model and f_θ is the original model. Please refer to Appendix E for a more detailed description.

Results From Table 3, we can observe that the proposed re-weighing strategy effectively enhances the knowledge learning capability of the PEFT method in fine-tuning models. Regardless of model type or parameter size, the improvement in knowledge learning accuracy ranges from 3 to 5 points, indicating that this method this method can guide the model towards a more accurate direction of target knowledge. Moreover, except for a few cases, there is also an improvement in the the generality and locality of knowledge learning, demonstrating the method's effectiveness in guiding the model to utilize knowledge. Regarding models with different parameter sizes, it can be observed that the re-weighing strategy performs more prominently on smaller-scale models. This might be because models with fewer parameters are more susceptible to the guidance of the loss function.

4 Conclusion

In this paper, we adopt a semantic perspective to scrutinize PEFT's performance in guiding model knowledge learning. Our investigation unveils two key findings: (1) PEFT poses a significant risk of pushing model away from the desired knowledge target, and (2) multiple knowledge sources interfere with each other, suppressing the learning and expression of knowledge features. Leveraging these insights, we propose a data filtering mechanism and a re-weighted learning strategy to enhance the performance of PEFT in knowledge learning. Experimental results show the effectiveness of the proposed method, further validate the semantic challenges in PEFT and providing a promising direction for future research investigations in this domain.

Limitations

Despite the valuable insights from a semantic perspective into PEFT of model knowledge learning, our study has several limitations. First, although the proposed method can achieve a notable improvement in knowledge learning accuracy, the model's accuracy in multiple knowledge learning remains relatively low and requires further improvement. Secondly, our work primarily focus on the learning of factual knowledge, lacking exploration into other types of knowledge, and we prioritize the accuracy of knowledge learning while paying less attention to other aspects such as knowledge reasoning abilities. Third, due to hardware constraints, we primarily investigated models up to a scale of 7 billion parameters. Further research that replicates our study using larger-scale models would be beneficial in confirming our findings and refining the analysis proposed in our investigation from a mathematical perspective.

Ethics Statement

This research aims to ensure that the model can accurately learn knowledge. However, it's crucial to remember that harmful knowledge data could potentially result in the model generating harmful outputs. All of our data has undergone thorough human scrutiny, with any malicious knowledge or offensive content meticulously removed. The research outlined in this paper adheres strictly to ethical guidelines and principles. We have placed a high priority on privacy, addressed biases, ensured transparency, and advocated for responsible usage.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.62106105), the CCF-Baidu Open Fund (No.CCF-Baidu202307), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation. *CoRR*, abs/2309.14402.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 8493–8502. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5937–5947. Association for Computational Linguistics.

Saad Hikmat Haji and Adnan Mohsin Abdulazeez. 2021. Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4):2715–2743.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022. Aging with GRACE: lifelong model editing with discrete key-value adaptors. *CoRR*, abs/2211.11031.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023a. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023b. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. 2022. Plug-and-play adaptation for continuously-updated QA. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 438–447. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1774–1793. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023b. PMET: precise model editing in a transformer. *CoRR*, abs/2308.08742.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *CoRR*, abs/2303.15647.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Tianjiao Zhao, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*, 2305.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kaifeng Lv, Shunhua Jiang, and Jian Li. 2017. Learning gradient descent: Better generalization and longer horizons. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2247–2255. PMLR.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2833–2861. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Túlio Ribeiro. 2022. Fixing model bugs with natural language patches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11600–11613. Association for Computational Linguistics.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-source multi-type knowledge exploration and exploitation for dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12522–12537. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *CoRR*, abs/2312.05934.

George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. 2023. Empirical analysis of the strengths and weaknesses of PEFT techniques for llms. *CoRR*, abs/2304.14999.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Renzhi Wang, Jing Li, and Piji Li. 2023b. Infodiffusion: Information entropy aware diffusion process for non-autoregressive text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13757–13770. Association for Computational Linguistics.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023c. Resolving knowledge conflicts in large language models. *CoRR*, abs/2310.00935.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *CoRR*, abs/2312.12148.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10222–10240. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.

A Related Work

The existing literature on methods of knowledge acquisition can be broadly categorized into two streams based on whether it alters the parameters of the original model (Yao et al., 2023).

A.1 Preserve models' parameters

Retrieve augmentation This approach depends on an external knowledge base containing new or correct knowledge. The new knowledge base is seamlessly integrated with the base model, facilitating the retrieval of pertinent information in response to prompts (Murty et al., 2022; Madaan et al., 2022; Li et al., 2023a). For example, IKE (Zheng et al., 2023) employs an in-context learning approach to adjust LLMs outputs using demonstrations sourced from the corpus guided by similarity, thus circumventing the need for gradient calculations.

Adding additional parameters This paradigm introduces extra trainable parameters which represent new knowledge to LLMs while the original parameters keeping frozen. T-Patcher (Huang et al., 2023b) and CaliNET (Dong et al., 2022) both integrate neurons or patches into the last layer of the Feed-Forward Network (FFN), with T-Patcher employing one neuron per mistake and CaliNET incorporating multiple neurons for various knowledge cases. Conversely, GRACE (Hartvigsen et al., 2022) utilizes a discrete codebook as an Adapter to add and update elements over time, allowing for the modification of a model's predictions.

A.2 Modify models' parameters

Located and edit This approach initially identifies parameters linked to specific knowledge and adjusts them directly. The Knowledge Neuron (KN) technique (Dai et al., 2022) introduces a method for attributing knowledge to pinpoint the "knowledge neuron" (a crucial element within the FFN matrix) and then updates these neurons accordingly. ROME (Meng et al., 2022) employs causal mediation analysis to pinpoint the area requiring modification. Unlike KN, ROME doesn't focus

solely on altering knowledge neurons but instead modifies the entire matrix. However, both KN and ROME are limited to editing one factual association at a time. To address this limitation, MEMIT (Meng et al., 2023) builds upon ROME's framework, allowing for simultaneous editing across multiple cases. Building on MEMIT, PMET (Li et al., 2023b) incorporates attention values to achieve enhanced performance.

Parameter-efficient fine-tuning This method stands out as a widely embraced approach in the era of large-scale models, consistently yielding promising results across various downstream tasks. Recent advancements have introduced a series of PEFT techniques, such as Prefix-tuning (Li and Liang, 2021), Adapter-tuning (Houlsby et al., 2019), Prompt-tuning (Lester et al., 2021), $(IA)^3$ (Liu et al., 2022) and LoRA (Hu et al., 2022), which further enhance the appeal of knowledge learning through fine-tuning. AdaLoRA (Zhang et al., 2023) propose incremental parameter updates based on weight matrix importance assessment to enhance update efficiency and adaptability. Similarly, Plug-and-Play Adaptation (Lee et al., 2022) leverage regularized fine-tuning to enable large-scale continual learning for knowledge updating. **This paper focuses on the analysis of the parameter-efficient fine-tuning methods**.

B Dataset Details

ZsRE (Levy et al., 2017) is a Question Answering (QA) dataset using question rephrasing generated by back-translation as the equivalence neighborhood. **CounterFact** (Meng et al., 2022) is a more challenging dataset that accounts for counterfacts that start with low scores in comparison to correct facts. It constructs out-of-scope data by substituting the subject entity for a proximate subject entity sharing a predicate. This adjustment allows us to distinguish between superficial alterations in wording and more substantial modifications that align with a meaningful learning of knowledge. We follow previous data split (Cao et al., 2021; Meng et al., 2022; Mitchell et al., 2022a; Yao et al., 2023) to evaluate all the models on the test set. Following prior work (Mitchell et al., 2022a,b), we use the Natural Questions (NQ; Kwiatkowski et al. (2019)) as out-of-scope data to evaluate locality. We also adopt the extended version of ZsRE proposed by (Yao et al., 2023), which introduces a portability test for the original dataset.

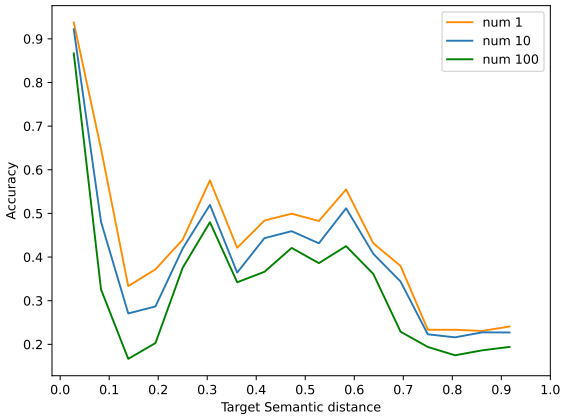

Figure 8: Results of learning various quantities of knowledge.

C Implementing Details

In most cases, we select the LLAMA2-7B-chat model as the subject of experimentation and fine-tuned the model using LoRA , a parameter adaptation technique widely employed during the era of large models. All model evaluations are conducted in zero-shot mode. To ensure the uniqueness of the model output, we set the temperature of the model to 0 during testing. The majority of fine-tuning experiments are conducted on NVIDIA RTX 3090 Tensor Core GPUs, while full fine-tuning is performed on NVIDIA A100 GPUs.

D Experiment on Varied Knowledge Quantities

In §2.1, We demonstrated that regardless of model type and parameter sizes, knowledge learning performance degradation occurs when the model learns knowledge of both closer and farther semantic distances. To further rigorously illustrate the universality of this phenomenon, we conducted experiments on models learning varying amounts of knowledge each iteration and found that the phenomenon persisted.

We utilized the ZsRE dataset and fine-tuned the LLaMA2-7B-chat model using LoRA. The relationship between the accuracy of knowledge learning by the model and the target semantic distance is illustrated in Figure 8. It can be observed that when learning different amounts of knowledge, the model's accuracy in knowledge learning consistently experiences significant declines at the same positions, thus validating the universality of the this phenomenon.

Model	ZsRE (LoRA)			COUNTERFACT (AdaLoRA)		
	Accuracy↑	Generality↑	Locality↑	Accuracy↑	Generality↑	Locality↑
BLOOMZ-1.7B	18.75	19.25	41.88	20.98	21.05	39.13
BLOOMZ-3B	15.60	16.52	42.15	18.48	17.34	43.47
BLOOMZ-7B	13.81	14.57	45.93	12.45	14.31	40.77
Mistral-7B	22.23	22.94	52.38	19.86	22.27	37.27
LLaMA2-7B	19.54	21.24	51.86	16.79	19.89	38.11

Table 4: The original result of the model. We follow the same experimental settings as §2.2.

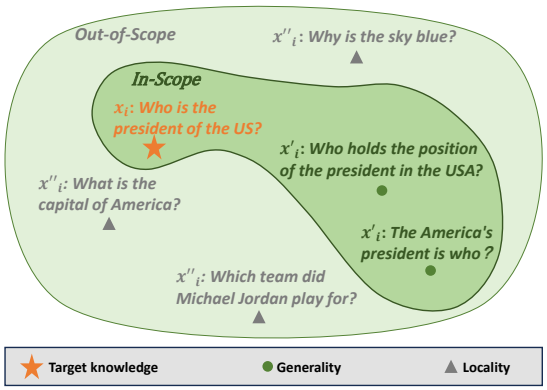

Figure 9: Explanation of the generality and locality for knowledge: "Who is the president of the US?" in a hypothetical semantic embedding space.

E Metric of Knowledge Learning

We aim to assess the model's learning of knowledge from a broader perspective. Following (Yao et al., 2023; Mitchell et al., 2022b), we employ the concepts of generality and locality to further evaluate the performance of knowledge learning.

Generality refers to the expectation that a fine-tuned model f_{θ^*} should also adjust neighboring equivalents $N(x_i, y_i^t)$ (in-scope in Figure 9), such as rephrased sentences. For example, as shown in Figure 9, the output of the question "The American's president is who?" and "Who holds the position of the president in the USA?" also need to be updated from "Donald Trump" to "Joe Biden". This is evaluated by the average accuracy of the model f_{θ^*} on examples uniformly sampled from the equivalence neighborhood:

$$\mathbb{E}_{(x'_i,y'_i)\sim N(x_i,y_i)} \mathbb{1}\left\{\text{argmax}_y f_{\theta^*}\left(y \mid x'_i\right) = y'_i\right\} \tag{9}$$

Locality dictates that updates should be applied locally, ensuring that the fine-tuned model does not alter the outputs of irrelevant examples $O(x_i, y_i)$. As shown in Figure 9, the output of the question "What is the capital of American" and "Which team did Michael Jordan play for?" are to be kept unchanged as the original model's output. Thus, locality is assessed by the consistency of predictions between the fine-tuned model f_{θ^*} and the original model f_θ:

$$\mathbb{E}_{(x'_i,y'_i)\sim O(x_i,y_i)} \mathbb{1}\left\{f_{\theta^*}\left(y \mid x'_i\right) = f_\theta\left(y \mid x'_i\right)\right\} \tag{10}$$

F Original Model Results

We provide a clearer report of the model's original performance on the knowledge dataset in Figure 4. It can be observed that the knowledge accuracy of fine-tuning the model based on LoRA is relatively low. Furthermore, the knowledge accuracy decreases as the model parameter increases. This may be due to the knowledge conflicts (Wang et al., 2023c), as more parameters imply a greater amount of internal knowledge held by the model, thereby increasing the probability of encountering knowledge conflicts.

We further reported some output results of the model. In this portion of the experiment, we used 100 instances of knowledge each time to fine-tune model. Due to the interference between knowledge, not only does it suppress the learning and expression of knowledge features (§2.2), but it also exacerbates the phenomenon of deviating from the target knowledge.

Target Semantic Distance	Model	Prompts and Target Knowledge	Model Response
Short	BLOOMZ-1.7B	Prompt: Which woman was the sister of Maria Elizabetha Jacson? Target Knowledge: Maria Theresa Jacson	Raw Response: Maria Jacson Fine-tuned Response: Anna Jacson
	BLOOMZ-3B	Prompt: What is the name of the chromosome where you can find RSPH6A? Target Knowledge: chromosome 19	Raw Response: chromosome 8 Fine-tuned Response: chromosome 6
	BLOOMZ-7B	Prompt: What day was USA-199 launched? Target Knowledge: 20 December 2007	Raw Response: December 15, 2011 Fine-tuned Response: December 12, 2011
	Mistral-7B	Prompt: What day was USA-199 launched? Target Knowledge: 20 December 2007	Raw Response: December 12 1995 Fine-tuned Response: 21 march 1997
	LLaMA2-7B	Prompt: What day was USA-35 launched? Target Knowledge: April 12, 1963	Raw Response: 9 April 1935 Fine-tuned Response: April 12, 1962
Medium	BLOOMZ-1.7B	Prompt: Who is listed as Wang Jipeng father? Target Knowledge: Wang Chonghua	Raw Response: Wang Jipeng Fine-tuned Response: Wang Jing
	BLOOMZ-3B	Prompt: Who is listed as Wang Jipeng father? Target Knowledge: Wang Chonghua	Raw Response: Wang jipeng Fine-tuned Response: Wang Jiping
	BLOOMZ-7B	Prompt: The inventor of Penrose stairs was whom? Target Knowledge: Richard Penrose	Raw Response: John Penrose Fine-tuned Response: Charles Babbage
	Mistral-7B	Prompt: Over which river does Dexter Coffin Bridge cross? Target Knowledge: Connecticut Creek	Raw Response: Saco River Fine-tuned Response: Connecticut dexter
	LLaMA2-7B	Prompt: During which historic war was Milton F. Pavlic an officer? Target Knowledge: Vietnam War	Raw Response: World War II Fine-tuned Response: Vietnam War
Long	BLOOMZ-1.7B	Prompt: What species is ZIC3 specific to? Target Knowledge: male	Raw Response: elegans Fine-tuned Response: drosophila melanogaster
	BLOOMZ-3B	Prompt: What species is ZIC3 specific to? Target Knowledge: male	Raw Response: human Fine-tuned Response: Zebrafish
	BLOOMZ-7B	Prompt: What species is ZIC3 specific to?, Target Knowledge: male	Raw Response: mammals Fine-tuned Response: human
	Mistral-7B	Prompt: Which was the network that originally hosted Jay Leno's Garage? Target Knowledge: NBC	Raw Response: Comedy Central Fine-tuned Response: CBS
	LLaMA2-7B	Prompt: At what location did John Walter Scott die? Target Knowledge: Windsor, Ontario, Canada	Raw Response: San Diego Fine-tuned Response: New York City

Table 5: Response of various models fine-tuned using LoRA on ZsRE dataset. The short, medium, and long target semantic distances correspond to semantic ranges of (0.1~0.3, 0.3~0.6, 0.6~1.0), respectively. The reason of choosing different prompts under identical target semantic distances across various models is that with the same prompt and target knowledge, the output knowledge of different models differs, and the parameters of word embedding within the models are different, leading to different target semantic distances as well.