Graceful Forgetting in Generative Language Models

Anonymous ACL submission

Abstract

Recently, the pretrain-finetune paradigm has become a cornerstone in various deep learning areas. While generally, the pre-trained model would provide both effectiveness and efficiency to downstream fine-tuning, studies have shown that not all knowledge acquired during pre-training is beneficial. Some of the knowledge may actually bring detrimental effects. To address this negative transfer problem, *graceful forgetting* has emerged as a promising approach. The core principle of graceful forgetting is to enhance the learning plasticity of the target task by selectively discarding knowledge from irrelevant tasks. However, this approach remains underexplored in the context of generative language models, and it is often challenging to migrate existing graceful forgetting algorithms to these models due to architecture incompatibility. To bridge this gap, in this paper we propose a novel framework, Learning With Forgetting (LWF), to achieve graceful forgetting in generative language models. With Fisher Information Matrix weighting the intended parameter updates, LWF computes forgetting confidence to evaluate self-generated knowledge regarding the forgetting task, and consequently, knowledge with high confidence is periodically unlearned during fine-tuning. We evaluate our framework on domain-specific question-answering tasks, demonstrating that, although determining the inter-task interaction mechanisms is still highly tricky, graceful forgetting can indeed lead to improved fine-tuning.

1 Introduction

In recent years, the *pretrain-finetune* paradigm has emerged as a dominant framework across natural language processing (NLP) tasks and various other domains (Zhou et al., 2023). This approach involves pre-training neural networks on large-scale corpora and subsequently fine-tuning on smaller, task-specific datasets to adapt to downstream applications. Its effectiveness has been evidenced by the success of prominent pre-trained models such as BERT (Devlin et al., 2019), GPT (Brown et al., 2020), and T5 (Raffel et al., 2020). These models have become the backbone of many state-of-the-art AI systems (OpenAI, 2023; Rombach et al., 2022), significantly fostering community development.

Despite tempting benefits such as data efficiency and reusability of this well-established paradigm, a long-standing issue, known as *negative transfer*, arises in exploiting pre-trained knowledge, referring to the fact that not all pre-trained knowledge contributes positively to the fine-tuning process of target learning tasks; in fact, some may even impair the learning plasticity of the target task (Zhang et al., 2023).

Negative transfer highlights a critical limitation of vanilla fine-tuning: it treats all pre-trained knowledge indiscriminately, which is not always the optimal practice. Inspired by this insight, many works were dedicated to developing more effective fine-tuning frameworks through alleviating negative transfer. Among these, a particularly promising approach is *graceful forgetting* (Wang et al., 2021; Liang and Li, 2023; Karakida and Akaho, 2022; Abbasi et al., 2024). Graceful forgetting, also known as active forgetting, is a concept originating in neuroscience, describing a memory mechanism in biological intelligence where the ability to acquire new knowledge is enhanced by selectively discarding irrelevant or outdated information (Anderson and Hulbert, 2021). Recent studies have shown that incorporating a similar "forgetting" mechanism into machine learning can also enhance the learning plasticity of new tasks (Zhou et al., 2022; D'Oro et al., 2023).

To achieve graceful forgetting, a variety of approaches have been explored. Some works focus on reflecting task synergy by exploiting metrics like gradient projections and loss interference, thereby reformulating gradient updates (Liang and

Li, 2023; Riemer et al., 2019). Alternatively, methods like BSS (Chen et al., 2019) and SRS (Shen et al., 2024) use structural regulation to implicitly forget pre-trained knowledge.

However, most existing methods are either tailored for vision tasks (particularly image classification) or are incompatible with pre-trained models, making it arduous to migrate them to pre-trained language models. Furthermore, the diverse and comprehensive information inherent in natural language data often results in ambiguous knowledge boundaries between different tasks, complicating the identification of explicit and granular inter-task correlations (Pruksachatkun et al., 2020). This issue is particularly profound in the rapidly evolving field of generative language models, where research on negative transfer and graceful forgetting remains extremely scarce.

To address this gap, in this paper we investigate the graceful forgetting in generative language models. The central question we try to explore is: *can generative language models achieve more effective fine-tuning by gracefully forgetting some unnecessary knowledge?* To answer this question, we propose a framework called Learning With Forgetting (**LWF**), to implement graceful forgetting in generative language models. Beginning from addressing the inaccessibility of pre-trained data, LWF leverages the nature of generative models, expressing knowledge regarding the forgetting task through self-generated texts. Then, given that identifying task-wise correlation is quite arduous, LWF calculates data-wise *forgetting confidence* for each data point by weighting the intended parameter updates with the Fisher Information Matrix. Finally, based on this confidence metric, LWF selects high-confidence data points and integrates machine unlearning techniques to periodically remove corresponding knowledge during the fine-tuning process. Experiments on domain-specific question-answering tasks demonstrate the superiorness of LWF over vanilla fine-tuning.

To the best of our knowledge, LWF represents the first systematic investigation of graceful forgetting in generative language models. Through extensive experiments and analyses, we validate the feasibility of promoting fine-tuning performance through graceful forgetting. Furthermore, our empirical findings shed light on some insights into this topic and hopefully offer inspiration for future investigation and innovation.

2 Related Work

2.1 Negative Transfer

The precise interpretation of *negative transfer* kindly varies among different research domains. In Multi-Task Learning (MTL), negative transfer refers to the performance degradation that occurs when learning conflicting tasks simultaneously (Go et al., 2023). Since the primary goal of MTL is to improve performance across all tasks, graceful forgetting is not a widely adopted strategy. Instead, methods to mitigate negative transfer in MTL typically focus on designing suitable criteria, such as gradient directions (Jiang et al., 2023) or signal-to-noise ratio (Go et al., 2023), to quantify the inter-task synergy and subsequently divide tasks into separate clusters.

In Continual Learning, negative transfer is regarded as a sacrifice of plasticity when pursuing stability (Karakida and Akaho, 2022). Specifically, the emphasis on retaining knowledge from previous tasks may result in reduced performance on learning new tasks. In this context, forgetting is employed as a counterbalance component to algorithms that overly prioritize memorizing past tasks, rather than as an independent mechanism to enhance performance on new tasks (Liang and Li, 2023; Wang et al., 2021; Schwarz et al., 2018).

Finally, as pre-trained models gradually become the critical foundation in various applications, research of negative transfer within the *pretrain-finetune* paradigm has gained increasing attention. This is the most consistent scenario with our work. A distinguishing characteristic of this paradigm is that pre-training involves training the model on vast amounts of data across diverse tasks, which is often inaccessible during fine-tuning. Consequently, current approaches to mitigating negative transfer in this context tend to rely on implicit forgetting methods, such as the Batch Spectral Shrinkage (Chen et al., 2019) and the Stable Rank Shrinkage (Shen et al., 2024).

2.2 Graceful Forgetting

Although forgetting is commonly regarded as an undesirable trait, suggesting the incapability of memorizing knowledge, recent studies have argued that striving for an omniscient model may be impractical due to limited model capacity and inevitable knowledge conflicts (Zhou et al., 2022; D'Oro et al., 2023). Inspired by neuroscience (Anderson and Hulbert, 2021), an increasing amount of research has explored the potential of improving learning plasticity through actively forgetting irrelevant knowledge. However, a universally accepted algorithm has yet to emerge, as the implementation is highly dependent on specific model architectures and task characteristics. Hitherto, most work in this area has focused on vision tasks, particularly image classification (Abbasi et al., 2024; Go et al., 2023; Wang et al., 2019; Liang and Li, 2023). For language tasks, (Chen et al., 2023) proposed a forgetting method that periodically resets the learned embedding layer to enhance the multi-lingual learning ability. But it is deployed during pre-training rather than fine-tuning. BSS (Chen et al., 2019) integrated their method into BERT and evaluated it on text classification tasks. Despite these trials, existing methods either lack generalizability or perform inadequately when applied to generative language models.

2.3 Machine Unlearning

Machine unlearning is a vibrantly investigated topic studying removing specific data, patterns, or knowledge from trained models (Yao et al., 2023). In the context of generative language models, unlearning is usually applied to aligning language models with human values, such as protecting user privacy (Patil et al., 2024), removing harmful contents (Liu et al., 2024), and reducing hallucination (Yao et al., 2023). A variety of unlearning strategies have been proposed for generative language models, including gradient ascent (Jang et al., 2023), localization-informed unlearning (Jang et al., 2023), influence function-based methods (Jia et al., 2024), and so on. While current generative language model unlearning research primarily focuses on eradicating undesirable behaviors, our work repurposes unlearning as a mechanism to achieve graceful forgetting, thereby enhancing the plasticity of fine-tuning. In essence, we leverage unlearning for better learning.

3 Methodology

In this section, we detail the implementation of our framework for graceful forgetting in generative language models. The framework primarily consists of three components: eliciting self-knowledge, evaluating forgetting confidence, and periodically unlearning. Fig 1 illustrates the overview. For the sake of convenience in exposition, we use $\mathcal{D}_L$ to represent the learning task and $\mathcal{D}_F$ to denote the forgetting task. But it is worth noting that the framework is task quantity-agnostic, which will be elaborated on Section 4.

3.1 Eliciting Self-Knowledge

The first step in forgetting specific knowledge is to acquire its representation. However, as discussed in Sec 2.1, the pre-training corpus is typically inaccessible in practice, making it uncertain whether $\mathcal{D}_F$ can adequately represent the model's knowledge. Fortunately, the inherent characteristics of generative language models provide a viable alternative: leveraging self-generated data. Specifically, we input the prompts (e.g., questions or instructions) from $\mathcal{D}_F$ into the base model $\mathcal{M}_{base}$ and collect its responses to form the unlearning dataset, which we donate as $\mathcal{D}_{self}$.

3.2 Evaluating Forgetting Confidence

Apparently, not all instances of forgetting are graceful. To ensure that forgetting enhances rather than hinders fine-tuning, we need a metric to reflect the confidence that forgetting specific knowledge will not lead to catastrophe. Moreover, given the rich and diverse semantic information present in natural language sentences, we argue that a task-level metric is too coarse-grained. Instead, we define forgetting confidence at the data point level, allowing for a more granular evaluation of which knowledge should be forgotten.

For a generated text x in $\mathcal{D}_{self}$, the posteriori $P(\mathcal{D}_L|x)$ intuitively reflects to what extent $\mathcal{D}_L$ and x are synergistic. The lower $P(\mathcal{D}_L|x)$ is, the more likely x is conflicted with $\mathcal{D}_{self}$. Considering $P(\mathcal{D}_L|x)$ is computationally intractable, we use $P(\mathcal{D}_L|\theta^*(x))$ as a surrogate, where

$$\theta^*(x) = \arg\max_{\theta} P(\theta|x) \tag{1}$$

Since only the relative value is required, we can use $P(\theta^*(x)|\mathcal{D}_L)$ to represent $P(\mathcal{D}_L|\theta^*(x))$, as the two are positively proportional according to the Bayes' Theorem. Based on this, we define the forgetting confidence as:

$$FC(x) \propto -\log P(\theta^*(x)|\mathcal{D}_L) \tag{2}$$

Following prior works (Kirkpatrick et al., 2017; Wang et al., 2021), we assume $P(\theta|\mathcal{D}_L)$ as a Gaussian distribution centered at $\theta^*_L = \arg\max_{\theta} P(\theta|\mathcal{D}_\mathcal{L})$, and this distribution can be approximated using a second-order Taylor expansion

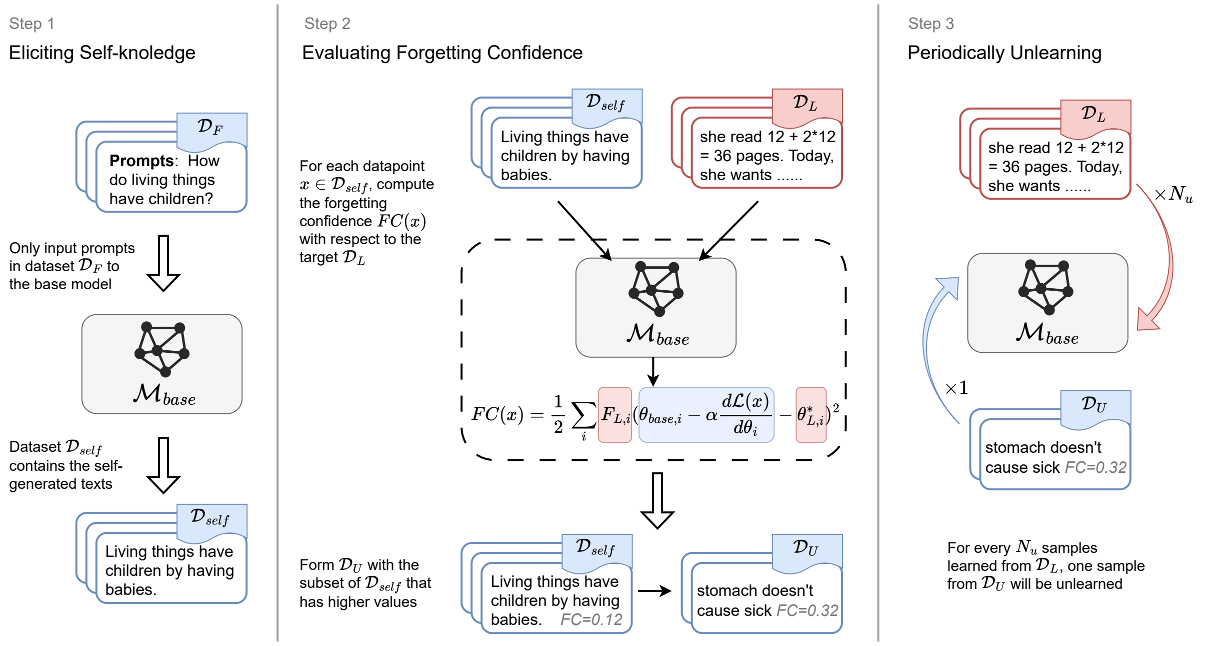

Figure 1: The overview of the LWF framework. Given the forgetting task $\mathcal{D}_F$ and learning task $\mathcal{D}_L$, LWF first constructs $\mathcal{D}_{self}$ through self-generated texts to represent the knowledge regarding the forgetting task. Then, with the Fisher Information Matrix F_L and the optimal parameters of the learning task approximated from $\mathcal{D}_L$, LWF calculates forgetting confidence for each data point in $\mathcal{D}_{self}$. Finally, data points with high forgetting confidence are selected for unlearning, represented by $\mathcal{D}_U$. The unlearning process is integrated into the fine-tuning process of $\mathcal{D}_L$ and is executed periodically at intervals of N_u.

around θ^*_L:

$$\log P(\theta|\mathcal{D}_L) \approx -\frac{1}{2}(\theta - \theta^*_L)^T \left(\frac{\partial^2 \log P(\theta|\mathcal{D}_L)}{\partial^2\theta}|_{\theta^*_L}\right)(\theta - \theta^*_L) \tag{3}$$

In practice, we integrate Equation 2 with Equation 3 and use a single-step update from the base model to represent $\theta^*(x)$, thereby reducing computational costs:

$$FC(x) = \frac{1}{2}\sum_i F_{L,i}(\theta_{base,i} - \alpha\frac{d\mathcal{L}(x)}{d\theta_i} - \theta^*_{L,i})^2 \tag{4}$$

F_L represents the Fisher Information Matrix (FIM), which is the negative expectation of the Hessian Matrix in Equation 3. The parameters of the base model are represented by θ_{base}, while $\mathcal{L}(x)$ refers to the cross-entropy loss of x. α controls the length of the single-step update. The θ^*_L is obtained by training the base model on $\mathcal{D}_L$.

Intuitively, Equation 4 measures the conflict between x and $\mathcal{D}_L$ by evaluating the alignment between the intended parameter update induced by x and the target θ^*_L. The FIM F_L serves as a weighting mechanism that captures the relative importance of each parameter.

3.3 Periodically Unlearning

With the forgetting confidence $FC(x)$, we can filter data from $\mathcal{D}_{self}$ with higher values to construct a subset $\mathcal{D}_U$. However, given the well-known "instability" characteristic of machine unlearning (Yao et al., 2023; Liu et al., 2024), casually unlearning $\mathcal{D}_U$ is unlikely to yield consistent improvements. Especially the data selected for forgetting in $\mathcal{D}_U$ is only potentially "conflicting" with the target task, rather than definitively harmful.

Therefore, to make the training process stable, we adopt a "periodically unlearning" strategy. Specifically, we perform learning on $\mathcal{D}_L$ and unlearning on $\mathcal{D}_U$ simultaneously, while maintaining a fixed interval N_u between consecutive unlearning sessions. For example, if the interval $N_u = 7$, then for every 7 data points learned from $\mathcal{D}_L$, one data point from $\mathcal{D}_U$ will be unlearned.

We use *Gradient Ascent* as our unlearning algorithm, which merely involves negating the loss function. For a periodic batch $\mathcal{X} = \{x^l_1, \ldots, x^l_{N_u}, x^u\}$ where $\{x^l_1, \ldots, x^l_{N_u}\} \subset \mathcal{D}_L$ and $x^u \in \mathcal{D}_U$, the loss can be written as:

$$\mathcal{L}_{pu}(\mathcal{X}) = \sum_{x \in \{x^l_1, \ldots, x^l_{N_u}\}} \mathcal{L}(x) - \beta\mathcal{L}(x^u) \tag{5}$$

where $\mathcal{L}$ is the *sft* loss and β is the unlearning rate.

4 Experiments

4.1 Setup

4.1.1 Datasets

To discuss the application and effectiveness of LWF, we deploy our method to domain-specific question-answering tasks. We select five datasets, each representing distinct domains of knowledge, to observe the outcomes of various learning-forgetting combinations. The datasets include: **gsm8k** (Cobbe et al., 2021) representing math; **qasc** (Khot et al., 2020) representing primary science; **sst5** (Socher et al., 2013) representing sentiment recognition; **dental**, the subset of MedMCQA (Pal et al., 2022) regarding dental knowledge; **psychol**, the subset of MMLU (Hendrycks et al., 2021) regarding psychology.

4.1.2 Implementation Details

We utilize Llama3.2-1B (Touvron et al., 2023) as the base model for all experiments. To elicit self-knowledge, we use a 3-shot prompt concatenated with the input question, employ a greedy decoding strategy, and constrain the maximum number of generated tokens to 256. When computing the forgetting confidence, we set the one step update coefficient α to 1e-2 (as defined in Equation 4). To maintain the coherence of batch gradient descent during periodically unlearning, we combine $\mathcal{D}_L$ and $\mathcal{D}_U$ by incorporating one data point from $\mathcal{D}_U$ for every N_u data points from $\mathcal{D}_L$. The training process uses a batch size of 4, a learning rate of 1e-5, and spans in total one epoch. N_u is set to 7, and β (as defined in Equation 5) is either 0.1 or 0.05, depending on forgetting tasks. All training procedures are carried out on eight NVIDIA RTX 4090 GPUs with full parameter tuning. For more details please refer to Appendix A.

4.2 Results on Question Answering

Table 1 shows the results on question-answering tasks. Each column represents different learning task and each row indicates different forgetting task. Specifically, the first row *none* means no forgetting task, that is, vanilla fine-tuning; and the last row *mixed* implies that the forgetting task comprises all datasets except the one used for learning.

As the results show, in most learning-forgetting combinations, LWF promotes the performance on the learning task compared to vanilla fine-tuning. Two exceptions are when learning dental with forgetting qasc and psychol. We believe this is attributed to the low forgetting confidence of self-generated samples in these two cases. The assumption is evidenced by the results of the last row, where mixing all other datasets as the forgetting task consistently improves performance on the learning task. This suggests that the distribution of forgetting confidence varies across datasets. Therefore, when the forgetting task is composed of a diverse set of datasets, the likelihood of improvement on the learning task grows, as there are potentially more high-confidence samples available for selection.

	gsm8k	qasc	sst5	dental	psychol
none	19.71	42.98	49.55	36.87	46.42
gsm8k	-	+4.03%	+2.83%	+1.46%	+6.33%
qasc	+5.38%	-	+2.54%	-4.53%	+5.54%
sst5	+2.67%	+3.02%	-	+0.22%	+0.41%
dental	+10.40%	+5.28%	+2.10%	-	+1.59%
psychol	+1.17%	+2.00%	+1.27%	-4.10%	-
mixed	+6.95%	+5.54%	+2.10%	+1.46%	+7.93%

Table 1: Results on domain-specific question answering. Each column shares the same learning task and the rows represent different forgetting tasks. All percentages are calculated based on *none*.

	qasc	sst5	dental	psychol
qasc	-	-14.93%	-0.37%	+1.00 %
sst5	-4.94%	-	-1.36%	-1.85%
dental	-1.45%	+0.17%	-	-3.65%
psychol	-17.43%	-12.02%	+4.80%	-

Table 2: Side-effect results. Percentages are calculated based on vanilla fine-tuning.

We also examine the side effects of LWF, *i.e.*, its impact on datasets that are neither part of the learning task nor the forgetting task (for simplicity we denote them as side-tasks). To evaluate, we compute the average accuracy on side-tasks and compare it to that of the vanilla fine-tuned model.

Table 2 shows the results, where each column represents the learning task and the row indicates the forgetting task. The gsm8k is excluded because its different format will cause superficial forgetting, which we discuss in Appendix B. As observed in the other tasks, the side effects vary depending on the specific learning-forgetting combinations. Overall, the impact is much milder when learning complex tasks like psychol and dental.

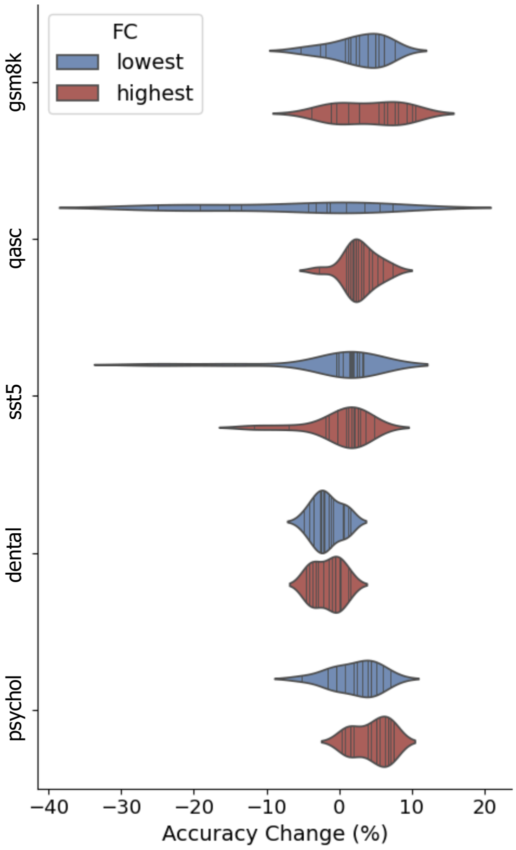

Figure 2: Distribution of accuracy changes between two filtering strategies. Percentages are calculated based on vanilla fine-tuning.

$\mathcal{D}_F$	gsm8k	qasc	sst5	dental	psychol
none	19.71	42.98	49.55	36.87	46.42
gsm8k	-	-65.6%	+0.5%	-0.8%	-9.9%
qasc	+5.0%	-	+0.4%	-7.4%	+3.6%
sst5	+4.3%	+7.8%	-	-3.5%	+9.1%
dental	-8.5%	-26.6%	+1.2%	-	+0.0%
psychol	-6.5%	-8.3%	-21.7%	-4.9%	-
mixed	-3.5%	-73.9%	+2.3%	+0.8%	-7.9%

Table 3: Results of *ahead unlearning*, ablation study for periodically unlearning, where unlearning is completed before fine-tuning.

$\mathcal{D}_F$	gsm8k	qasc	sst5	dental	psychol
none	19.71	42.98	49.55	36.87	46.42
gsm8k	-	-12.6%	-0.1%	-4.9%	+10.3%
qasc	-8.8%	-	+3.9%	-2.0%	+7.7%
sst5	-8.8%	-10.8%	-	-4.3%	+3.6%
dental	-6.5%	-6.9%	+1.8%	-	+7.5%
psychol	-5.0%	-13.1%	+2.1%	-2.7%	-
mixed	-6.5%	-10.1%	-0.5%	-4.9%	+9.5%

Table 4: Results of *randomly unlearning*, ablation study for periodically unlearning, where unlearning is randomly executed during fine-tuning.

4.3 Analysis on Forgetting Confidence

In LWF, one of the most critical steps is computing the forgetting confidence, which reflects the confidence that forgetting specific data will positively contribute to the learning task. To achieve this, we propose the computable metric $FC(x)$ (in Sec 3.2). However, the computation involves approximation errors. More importantly, the relation between the final model performance and to what extent the gradient update direction of individual data aligns with expectations is not a definitive mapping. These limitations make $FC(x)$ a non-absolute measurement. In practice, we also observed that unlearning data points with relatively lower FC does not absolutely result in degradation.

To better understand the role of forgetting confidence from a statistical perspective, we design a comparison for LWF that employs the opposite filtering strategy, *i.e.*, selecting data with the lowest forgetting confidence. For each learning task, we calculate the accuracy change percentage of these two strategies across all forgetting tasks, based on vanilla fine-tuning results. To enlarge the sample size, we collected results across four different unlearning rate, $\beta \in \{0.05, 0.10, 0.20, 0.25\}$.

Fig 2 shows the distribution of two strategies. The *red* part corresponds to unlearning data with the highest FC and the *blue* part represents unlearning data with the lowest. The x-axis is the accuracy change percentage. As we can see, generally the privilege of unlearning data with high FC manifests in two aspects. First, the average accuracy enhancement is higher. Second, the enhancement is more stable. Specifically, the variance and overall range are smaller when selecting data with high FC, while unlearning data with low FC may lead to highly variable results, including extremely poor cases. In conclusion, unlearning data with high forgetting confidence yields averagely better performance and ensures greater robustness, making it a more reliable approach in practice.

4.4 Abaltion on Periodically unlearning

To alleviate the vulnerability of machine unlearning, we propose the periodically unlearning strategy to stabilize the training process. In this section, we conduct an ablation study to demonstrate that periodically unlearning is the most suitable strategy for effectively combining the learning and unlearn-

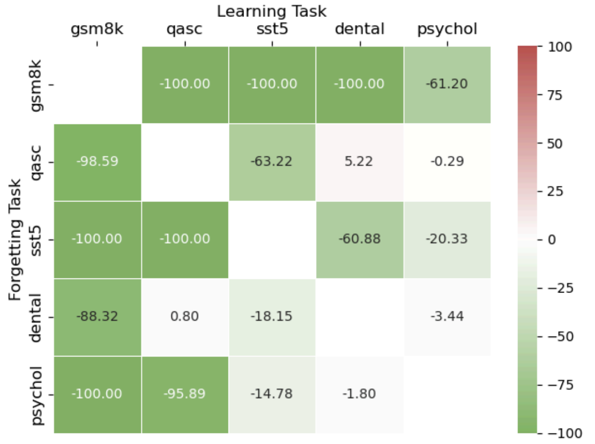

Figure 3: Accuracy change percentage of the forgetting task across different learning-forgetting combinations. Percentages are computed based on vanilla fine-tuning.

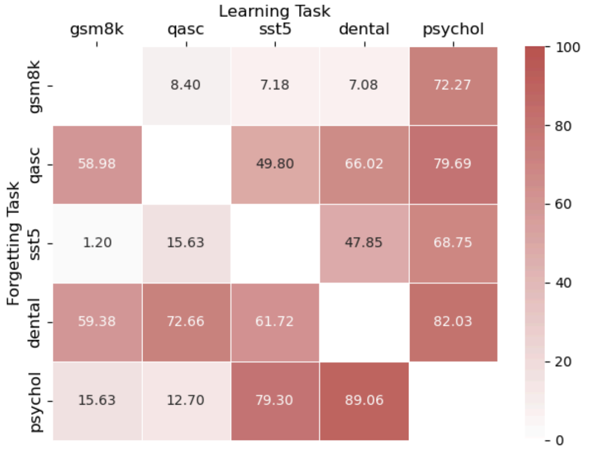

Figure 4: Cosine similarity between the outputs of forgetting tasks generated by the vanilla fine-tuned model and LWF resulting model. Values are multiplied by 100.

ing processes.

We design two variants of unlearning strategies for comparison. The first strategy involves conducting the unlearning process prior to the learning process, which we refer to as *ahead unlearning*. The second strategy allows the model to randomly execute unlearning during fine-tuning, which we name as *randomly unlearning*. It is important to note that, together with *periodically unlearning*, all three strategies share the same ratio of learning and unlearning samples.

Table 3 and Table 4 present the results of *ahead unlearning* and *randomly unlearning* respectively. As shown, both of them are significantly less effective than periodically unlearning (Table 1) in general. In both of them, the majority of learning-forgetting combinations lead to negative changes. Furthermore, *ahead unlearning* demonstrates extremely undesirable instances. This may be attributed to that conducting the unlearning process on the base model in advance may cause uncontrolled damage to the pre-trained knowledge. If critical and foundational knowledge is affected, the subsequent learning process may suffer severe degradation. Therefore, it can be concluded that mixing the learning and unlearning processes is better than conducting them separately, and switching them periodically is superior to randomly.

4.5 Analysis on the Forgotten Task

In this section, we turn our attention to the forgetting task, examining how the model's performance on it changes before and after the deployment of LWF. While it is intuitively reasonable that the

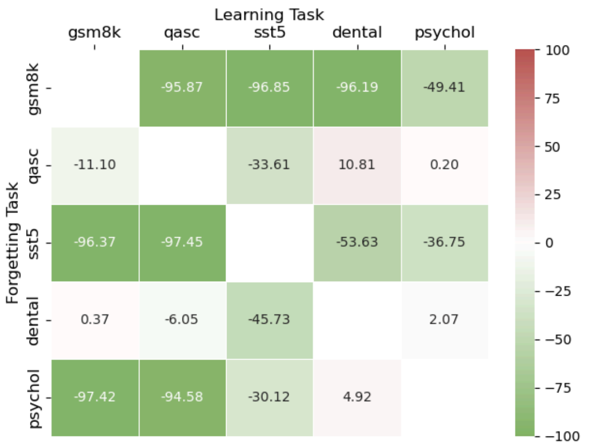

Figure 5: TTR change percentage of the forgetting task across different learning-forgetting combinations. Percentages are computed based on vanilla fine-tuning.

model's knowledge about the forgetting task will be nearly erased, the empirical findings are more complicated. Fig 3 is a heatmap illustrating the percentage changes in accuracy for the forgetting tasks compared to vanilla fine-tuning. As depicted, although the accuracy of the forgetting tasks generally declines after LWF across most learning-forgetting combinations, the extent varies largely. Approximately half of the cases exhibit a drop close to 100%, while others show only marginal decreases, particularly when either the learning task or the forgetting task involves *dental* or *psychol*.

Additionally, we noticed that the accuracy merely reflects whether the final answer is correct, which is insufficient to fully capture the nuanced changes in the generated outputs of generative models. Therefore, we take a further step to analyze semantic changes. Specifically, we use SimCSE model [1] (Gao et al., 2021) to obtain the sentence vectors of the responses generated by the vanilla fine-tuned model and LWF model for the same forgetting task question. We then compute the cosine similarity between these vectors to quantify semantic differences. The results are shown in Fig 4.

Considering a cosine similarity score above 80% is generally required to confidently assert that two sentences are semantically similar, the figure reveals that the semantic changes in most combinations are substantial. Notably, the cases where the similarity approaches or exceeds 80% align almost entirely with those exhibiting minimal accuracy drops, that is, combinations involving *dental* or *psychol*. We believe this phenomenon may be attributed to that *dental* and *psychol* are inherently more complex than the other tasks. Learning or forgetting these domains requires the model to engage with more sophisticated knowledge structures, making it less susceptible to extreme forgetting.

We also evaluate changes in lexical diversity, as shown in Fig 5, with Type Token Ratio (TTR) as the metric. Similar to the trends observed in accuracy changes, the TTR experiences a significant decline in most combinations, and the cases that maintained high correctness and semantic similarity also largely preserved their lexical diversity.

4.6 Multi-Task Learning

In this section, we discuss the performance of the LWF in multi-task learning scenarios. Specifically, we select one from the five datasets as the forgetting task and combine the remaining four to form the learning tasks. To mitigate the risk of catastrophic forgetting of earlier tasks, we evenly mix the learning tasks during training. Fig 6 shows the comparison of multi-task accuracy between LWF and vanilla fine-tuning. As we can see, LWF demonstrates an overall improvement in performance compared to vanilla fine-tuning, while at the individual task level, not all learning tasks benefit equally from LWF.

The results underscore the complexity of multi-task learning scenarios. Beyond the overall gains achieved through graceful forgetting, there are intricate interactions among the learning tasks themselves. Improvements in one task may inadvertently suppress the performance of others.

[1] https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased

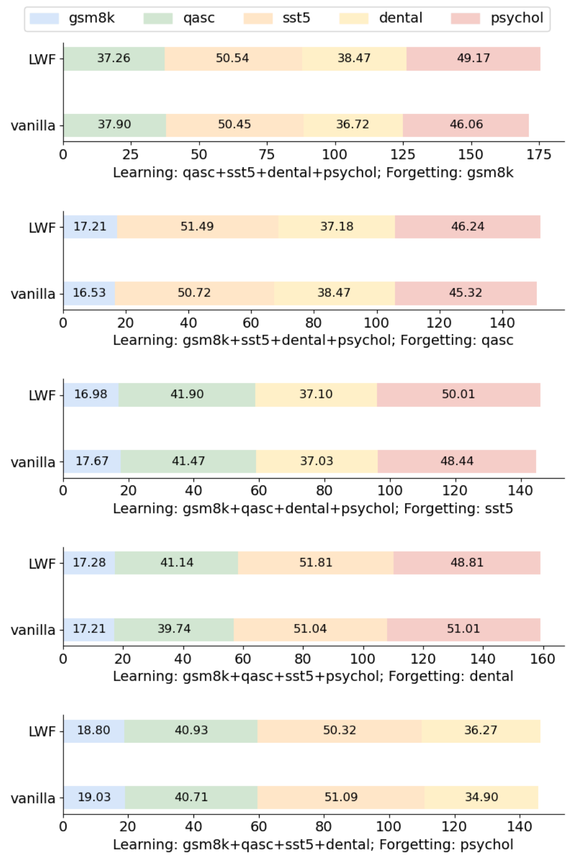

Figure 6: Accuracy results in the multi-task learning setting. Labeled below each subplot are the forgetting task and learning tasks.

5 Conclusion

In this paper, we propose a novel framework, Learning With Forgetting (LWF), to achieve graceful forgetting in generative language models. LWF addresses the inaccessibility of pre-trained data by leveraging self-generated knowledge, calculates forgetting confidence for each data point by weighting the intended parameter update with Fisher Information Matrix, and employs gradient ascent to periodically unlearn high-confidence data during fine-tuning. Empirical results on domain-specific question-answering tasks demonstrate the effectiveness of LWF. Furthermore, we conduct extensive experiments to analyze the contribution of each component of LWF, the effects of forgetting specific tasks, and the framework's performance in learning or forgetting multiple tasks. While fully elucidating the mechanisms of inter-task interactions and achieving optimal graceful forgetting still need great effort, we hope our work provides valuable insights into this emerging area and inspires future research endeavors.

Limitation

As the first framework for graceful forgetting in generative language models, LWF has several areas that warrant further improvement. The first area concerns the measurement of *forgetting confidence*. Quantifying the interactions among learning data has long been a challenge in the field of knowledge transfer. In this paper, we adopt a popular perspective of intended parameter updates. While empirical results demonstrate its statistical effectiveness in selecting better data and enhancing robustness, its applicability is likely to decrease when the available data for selection is limited. We anticipate that future advancements in knowledge transfer research will yield more precise and reliable metrics for measuring forgetting confidence.

The second area pertains to the unlearning process. While LWF demonstrates the feasibility of graceful forgetting through machine unlearning, the adopted algorithm, *gradient ascent*, is a relatively naive approach within machine unlearning algorithms. As observed, this method may introduce instability and unintended side effects. Although we have implemented countermeasures, such as periodically unlearning, to mitigate these issues, we believe that future work could benefit from the development of more sophisticated and tailored unlearning algorithms.

References

Ali Abbasi, Parsa Nooralinejad, Hamed Pirsiavash, and Soheil Kolouri. 2024. Brainwash: A poisoning attack to forget in continual learning. In *CVPR*, pages 24057–24066. IEEE.

Michael C Anderson and Justin C Hulbert. 2021. Active forgetting: Adaptation of memory by prefrontal control. *annual review of psychology*, 72(1):1–36.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, pages 1906–1916.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Lars Erik Saito Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. In *NeurIPS*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron C. Courville. 2023. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *ICLR*. OpenReview.net.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP (1)*, pages 6894–6910. Association for Computational Linguistics.

Hyojun Go, JinYoung Kim, Yunsung Lee, Seunghyun Lee, Shinhyeok Oh, Hyeongdon Moon, and Seungtaek Choi. 2023. Addressing negative transfer in diffusion models. In *NeurIPS*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *ACL (1)*, pages 14389–14408. Association for Computational Linguistics.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. SOUL: unlocking the power of second-order optimization for LLM unlearning. In *EMNLP*, pages 4276–4292. Association for Computational Linguistics.

Junguang Jiang, Baixu Chen, Junwei Pan, Ximei Wang, Dapeng Liu, Jie Jiang, and Mingsheng Long. 2023. Forkmerge: Mitigating negative transfer in auxiliary-task learning. In *NeurIPS*.

Ryo Karakida and Shotaro Akaho. 2022. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting. In *ICLR*. OpenReview.net.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *AAAI*, pages 8082–8090. AAAI Press.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yan-Shuo Liang and Wu-Jun Li. 2023. Adaptive plasticity improvement for continual learning. In *CVPR*, pages 7816–7825. IEEE.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *CoRR*, abs/2402.08787.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *CHIL*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *ICLR*. OpenReview.net.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *CoRR*, abs/2005.00628.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR (Poster)*. OpenReview.net.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4535–4544. PMLR.

Sicong Shen, Yang Zhou, Bingzheng Wei, Eric I-Chao Chang, and Yan Xu. 2024. Tuning stable rank shrinkage: Aiming at the overlooked structural risk in fine-tuning. In *CVPR*, pages 28474–28484. IEEE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642. ACL.

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. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. 2021. AFEC: active forgetting of negative transfer in continual learning. In *NeurIPS*, pages 22379–22391.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime G. Carbonell. 2019. Characterizing and avoiding negative transfer. In *CVPR*, pages 11293–11302. Computer Vision Foundation / IEEE.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2023. A survey on negative transfer. *IEEE CAA J. Autom. Sinica*, 10(2):305–329.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. A comprehensive survey on pretrained foundation models: A history from BERT to chatgpt. *CoRR*, abs/2302.09419.

Hattie Zhou, Ankit Vani, Hugo Larochelle, and Aaron C. Courville. 2022. Fortuitous forgetting in connectionist networks. In *ICLR*. OpenReview.net.

	EN	IT	ZH	ES	TR
none	19.71	6.67	9.78	7.81	9.10
EN	-	+6.90%	+2.35%	+3.84%	+4.18%
IT	+5.38%	-	-5.42%	+1.92%	-14.18%
ZH	+2.69%	-30.73%	-	-3.84%	-9.23%
ES	+0.41%	-35.23%	-6.24%	-	+5.05%
TR	+6.95%	-25.04%	-8.49%	-1.02%	-
mixed	+3.45%	+25.04%	+7.77%	+10.63%	+20.77%

Table 5: Results on multi-lingual question-answering.

	gsm8k	qasc	sst5	dental	psychol	AVG.
vanilla-FT	19.71	42.98	49.55	36.87	46.42	39.12
BSS	20.39	44.28	49.73	35.51	44.77	38.94
SRS	17.36	40.28	50.50	35.05	46.61	37.96
LWF-mixed	21.08	45.36	50.59	37.41	50.10	40.91

Table 6: Results of structural regulation methods

A More Implementation Details

Table 7, Table 8, Table 9, Table 10, and Table 11 present the few-shot CoT prompts designed for each dataset, which are utilized during both the self-knowledge elicitation and evaluation stages. As illustrated in the prompts, answers are formatted with the phrase '*The answer is*' to facilitate the convenience of extracting answers. Any model output that deviates from this format is considered incorrect. In cases where multiple occurrences of '*The answer is*' appear in the output, the first instance is treated as the definitive answer.

B Superficial Forgetting

Although all five datasets used in our experiments are question-answering datasets, gsm8k differs significantly in format from the other four. Specifically, gsm8k is free-form numerical question-answering, while the other four datasets are multiple-choice question-answering (see examples in Table 7 and Table 8). We observed that this format discrepancy can lead to significant side effects when gsm8k is the learning task and the other datasets are the forgetting tasks. By analyzing the model's outputs, we identified that the model trained under this setting often fails to generate answers in the multiple-choice format.

To illustrate, Table 12 provides examples of the model's output sentences on three tasks when gsm8k is the learning task and qasc is the forgetting task. As shown, while the rationale portion of the output appears coherent, the model fails to select a valid option at the end of its response.

This phenomenon suggests that, under the LWF framework, the model tends to focus on the most superficial pattern differences to distinguish the learning task from the forgetting task. Therefore, to mitigate extreme side effects, it is better to ensure that there are no overly superficial format differences between the learning and forgetting tasks.

C LWF in Multi-Lingual Tasks

Beyond domain-specific question-answering, another promising application of graceful forgetting is in multilingual learning, where hopefully the model's performance in a target language can be enhanced by forgetting its representations in other languages. To explore this, we applied LWF to the gsm8k dataset across five languages: English (EN), Italian (IT), Chinese (ZH), Spanish (ES), and Turkish (TR). The results are presented in Table 5.

We observed an interesting phenomenon: aside from the *mixed* approach, which consistently improves performance across all languages, the results are all positive when learning English with forgetting other languages, or vice versa. However, when the combination does not involve English, the outcomes are predominantly negative. We hypothesize that this phenomenon may stem from the disproportionate volume of English data used during the model's pre-training, which likely stabilizes LWF's performance when English is involved.

D Structural Regulation Methods

As discussed in Section 2.1, structural regulation methods such as BSS (Chen et al., 2019) and SRS (Shen et al., 2024) are typical approaches for mitigating negative transfer within the pretrain-finetune paradigm. However, both of them have only been evaluated on classification tasks. We attempted to adapt them for generative language models, and the results of our implementations are presented in Table 6, along with vanilla fine-tuning (vanilla-FT) and LWF (LWF-mixed).

As observed, the performance of BSS and SRS is consistently inferior to LWF-mixed. In some cases, these methods even result in negative changes compared to vanilla-FT. We attribute this ineffectiveness to architecture differences. Specifically, BSS and SRS rely on operations applied to feature vectors, whereas the hidden states in generative language models are not inherently designed to represent sentence embeddings, the models for which BSS and SRS were originally designed.

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? Let's think step by step.
Answer: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted 21 - 15 = 6 trees. The answer is 6.
Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot? Let's think step by step.
Answer: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.
Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total? Let's think step by step.
Answer: Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39.

Table 7: Few-shot prompts for gsm8k.

Question: What type of water formation is formed by clouds? (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid Let's think step by step.
Answer: Beads of water are formed by water vapor condensing. Clouds are made of water vapor. Beads of water can be formed by clouds. The answer is (F).
Question: Removing what from food will preserve it? (A) flavor (B) body water (C) heat energy (D) color (E) Water (F) Bodily water (G) moisture (H) ingredients Let's think step by step.
Answer: Dehydrating food is used for preserving food. Dehydration preserves foods by removing moisture. Removing moisture from food preserves it. The answer is (G).
Question: Reproduction is the process by which living things what? (A) Most plants (B) allow growth (C) spread flower seeds (D) have wide set eyes (E) members of their own species (F) have birthing hips (G) have quiet laughter (H) give birth to babies Let's think step by step.
Answer: Reproduction is the process by which living things give rise to offspring. Whenever it starts to give birth, it gives birth up to two to four babies offspring. Reproduction is the process by which living things give birth to babies. The answer is (H).

Table 8: Few-shot prompts for qasc.

Question: What is the emotional attitude reflected in the sentence: "in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically ."? (A) very negative (B) negative (C) neutral (D) positive (E) very positive Let's think step by step:
Answer: The phrase "first stab" suggests this is an initial attempt, which can imply inexperience but also experimentation. "Slightly anarchic approach" does have a mild negative connotation with "anarchic," but it's softened by "slightly." "Works only sporadically" indicates mixed results. It doesn't work consistently, but it does work sometimes, which isn't entirely negative. Given these considerations, while there are elements that lean towards a negative interpretation, the overall tone of the sentence is not strongly positive or negative. So the emotional attitude is neutral. The answer is (C)
Question: What is the emotional attitude reflected in the sentence: "one long string of cliches."? (A) very negative (B) negative (C) neutral (D) positive (E) very positive Let's think step by step:
Answer: The phrase "one long string" suggests an extended or continuous sequence, which in this context implies that there are many instances of something undesirable. The term "cliches" refers to expressions or elements that have been overused to the point of losing their original impact, freshness, or appeal. It generally carries a negative connotation because it indicates a lack of originality or creativity. Considering these points, the sentence does not contain any positive words or phrases and is clearly pointing out a flaw or criticism. Therefore, the emotional attitude reflected in the sentence is negative. Since it's a straightforward criticism without strong intensifiers (like "extremely" or "utterly"), we can conclude that the most fitting emotional attitude is negative. The answer is (B)
Question: What is the emotional attitude reflected in the sentence: "although laced with humor and a few fanciful touches, the film is a refreshingly serious look at young women."? (A) very negative (B) negative (C) neutral (D) positive (E) very positive Let's think step by step:
Answer: The word "although" introduces a concession, acknowledging that the film contains elements of humor and some imaginative aspects ("fanciful touches"). This part suggests that there are positive qualities present in the film. The main clause uses the term "refreshingly," which conveys a positive and invigorating experience. It indicates that the seriousness of the film's approach to its subject matter is not only appreciated but also seen as something new or different from what might be expected. The focus on "young women" being treated seriously adds depth to the evaluation. It implies that the film offers a thoughtful and mature perspective on its subject, which can be seen as commendable. Given these considerations, the overall tone of the sentence is appreciative and positive. The use of "refreshingly" especially signals a positive emotional response to the film's treatment of its subject. However, since the positivity is moderate and not overly enthusiastic or exuberant, the most fitting emotional attitude is positive. The answer is (D).

Table 9: Few-shot prompts for sst5.

Question: Resistance to lateral shifting or anteroposterior shifting is called as:? (A) Retention. (B) Stability. (C) Support. (D) None. Let's think step by step: **Answer:** The question is asking about the term used to describe resistance to lateral (side-to-side) or anteroposterior (front-to-back) shifting. Retention generally refers to the ability to hold or keep something in place. In dentistry, for example, it refers to the resistance of a dental prosthesis to displacement in a vertical direction. Stability refers to the resistance to movement, particularly lateral or anteroposterior shifting. It is the ability of an object or structure to maintain its position when subjected to forces. Support refers to the ability to bear weight or hold something up, but it does not specifically refer to resistance to shifting. Based on the definitions, stability is the term that specifically refers to resistance to lateral or anteroposterior shifting. The answer is (B). **Question:** Tannerella forsythia is a ? (A) Nonmotile, gram-negative obligate anaerobe (B) Motile, gram-negative obligate anaerobe (C) Nonmotile, gram-negative obligate aerobe (D) Motile, gram-negative obligate aerobe Let's think step by step: **Answer:** Tannerella forsythia is a bacterium associated with periodontal disease. It is known to thrive in environments with little to no oxygen. Tannerella forsythia is nonmotile, meaning it does not have structures like flagella to move on its own. It is a gram-negative bacterium, which means it has a thin peptidoglycan layer and an outer membrane. It is an obligate anaerobe, meaning it cannot survive in the presence of oxygen and requires an oxygen-free environment to grow. So, it is a nonmotile, gram-negative obligate anaerobe. The answer is (A). **Question:** Which of the following bacteria is not a primary colonizer:? (A) Streptococci. (B) Actinomyces. (C) F.nucleatum. (D) None. Let's think step by step: **Answer:** Primary colonizers are the first bacteria to adhere to the tooth surface and form dental plaque. They are typically gram-positive bacteria, such as Streptococci and Actinomyces. Fusobacterium nucleatum is not a primary colonizer. It is a secondary colonizer that adheres to the primary colonizers and helps bridge the gap between early and late colonizers. The answer is (C).

Table 10: Few-shot prompts for dental.

Question: Ani believes that her attitudes and behavior play a central role in what happens to her. Such a belief is likely to be associated with? (A) a strong superego. (B) low self-esteem. (C) low self-efficacy. (D) an internal locus of control. Let's think step by step: **Answer:** The question is asking about a psychological construct that aligns with the belief that one's own actions and attitudes influence life events. This is closely related to the concept of locus of control. Locus of control refers to the extent to which individuals believe they can control events that affect them. There are two types: 1.Internal locus of control: Individuals believe they have control over their lives through their actions and decisions; 2.External locus of control: Individuals believe that external forces (e.g., luck, fate, or other people) control their lives. Ani's belief that her attitudes and behavior play a central role in what happens to her indicates an internal locus of control. She feels responsible for her outcomes and believes her actions matter. The answer is (D). **Question:** According to Caplan's model of consultee-centered case consultation, the consultant is primarily interested in? (A) identifying the causes and solutions of the client's presenting problems (B) identifying and eliminating the causes of the consultee's difficulties in handling a problem (C) establishing a hierarchy of authority to enable effective decision making (D) presenting a single, well-defined and unambiguous course of action for the consultant to overcome skills deficits Let's think step by step: **Answer:** Caplan's model of consultee-centered case consultation focuses on helping the consultee (e.g., a teacher, therapist, or other professional) improve their ability to handle a specific case or problem. The consultant does not directly intervene with the client but instead works with the consultee to address their difficulties in managing the situation. The primary goal is to identify and address the consultee's difficulties, which may stem from a lack of knowledge, skills, confidence, or objectivity. The consultant helps the consultee overcome these issues so they can better handle the client's problem. The correct answer is (B), as Caplan's model is primarily concerned with identifying and addressing the consultee's difficulties in handling a problem. The answer is (B). **Question:** Pascale is interested in the processing strategies children use to learn new information. Pascale would best be classified as what type of psychologist? (A) sociocultural (B) clinical (C) cognitive (D) behaviorist Let's think step by step: **Answer:** The question is asking about the type of psychologist Pascale would be classified as, based on her interest in processing strategies and learning. This aligns with the field of psychology that studies mental processes such as thinking, memory, and learning. Sociocultural psychologists focus on how social and cultural factors influence behavior and mental processes. While this could involve learning, it is not primarily about individual cognitive processing strategies. Clinical psychologists focus on diagnosing and treating mental health disorders. This does not align with Pascale's interest in learning and processing strategies. Cognitive psychologists study mental processes such as thinking, memory, and learning. This directly aligns with Pascale's interest in how children process and learn new information. Behaviorists focus on observable behaviors and how they are influenced by the environment. While learning is a part of behaviorism, it does not emphasize internal processing strategies as cognitive psychology does. The answer is (C).

Table 11: Few-shot prompts for psychol.

Learning task: gsm8k **Forgetting task:** qasc
sst5: The phrase "no movement" suggests a lack of action or movement, which in this context implies a lack of interest or enthusiasm. The term "yuks" is a slang term for laughter, which in this context carries a negative connotation because it indicates a lack of seriousness or sincerity. The phrase "not much of anything" suggests a lack of substance or depth, which in this context implies a lack of complexity or sophistication. Considering these points, the sentence does not contain any positive words or phrases and is clearly pointing out a flaw or criticism. Therefore, the emotional attitude reflected in the sentence is negative. Since it's a straightforward criticism without strong intensifiers (like "extremely" or "utterly"), we can conclude that the most fitting emotional attitude is negative. The answer is (6 - 2 - 2) = 2 **dental:** Heavy forces on periodontal ligament cause osteoclastic activity around the tooth, which results in the formation of new bone around the tooth. This new bone helps to support the tooth and prevent further damage. The answer is (6 - 2 - 2) = 2 **psychol:** The heritability of a trait is the proportion of the trait's variation that is due to genetic factors. In a population with 100 individuals, the heritability of a trait is 100% - 25% = 75%. The answer is 75

Table 12: Cases of superficial forgetting.