Graceful Forgetting in Generative Language Models

Chunyang Jiang, Chi-min Chan, Yiyang Cai, Yulong Liu, Wei Xue✉, Yike Guo✉

Hong Kong University of Science and Technology

cjiangaq@connect.ust.hk

Abstract

Recently, the pretrain-finetune paradigm has become a cornerstone in various deep learning areas. While in general the pre-trained model would promote both effectiveness and efficiency of downstream tasks 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 the fine-tuning tasks, which is also known as *negative transfer*. To address this 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 irrelevant knowledge. However, this approach remains underexplored in the context of generative language models, and it is often challenging to migrate existing 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. Our experiments demonstrate that, although thoroughly uncovering the mechanisms of knowledge interaction remains challenging in pre-trained language models, applying graceful forgetting can contribute to enhanced fine-tuning performance.

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 the model on large-scale corpora and subsequently fine-tuning it 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). And these models have become the backbone of many state-of-the-art AI applications (OpenAI, 2023; Rombach et al., 2022).

Despite offering compelling benefits such as data efficiency and reusability, this well-established paradigm continues to face a long-standing and prevalent issue, *negative transfer*, which surfaces in a new guise. While the typical interpretation of negative transfer in transfer learning refers to the performance degradation when learning conflicting tasks simultaneously or sequentially (Zhang et al., 2023a), its manifestation in the context of the pretrain-finetune paradigm takes on a different form, *the negative contribution of some pre-trained knowledge to the target fine-tuning task* (Zhang et al., 2023b). This problem highlights a critical limitation of vanilla fine-tuning: treating all pre-trained knowledge indiscriminately is not always the optimal practice.

To address this issue, a promising approach is to suppress the influence of potentially harmful knowledge, a strategy known as *graceful forgetting* (or active forgetting). Originating in neuroscience, this concept describes a memory mechanism in biological intelligence where the ability to acquire new knowledge is enhanced by selective elimination of irrelevant or outdated information (Anderson and Hulbert, 2021). Recent advances have demonstrated the feasibility of emulating this mechanism in machine learning models (Zhou et al., 2022; D'Oro et al., 2023), leading to its adoption in various studies aimed at enhancing learning plasticity (Wang et al., 2021; Chen et al., 2023; Liang and Li, 2023; Shen et al., 2024).

However, most existing graceful forgetting meth-

---

ods are either tailored to vision tasks or designed for non-autoregressive models, making their efforts incompatible or less effective when migrating to generative language models. This discrepancy primarily stems from the ambiguous knowledge boundaries inherent in language generation, which significantly complicate the identification of explicit and granular inter-task correlations (Pruksachatkun et al., 2020).

To address this gap, in this paper we investigate the graceful forgetting in generative language models. The central question guiding our study is: *can generative language models achieve more effective fine-tuning by gracefully forgetting unnecessary knowledge?* To answer this question, we propose a framework called Learning with Forgetting (**LWF**) to enable graceful forgetting in generative language models. Beginning with addressing the inaccessibility of pre-trained data, LWF leverages the inherent capabilities of generative models by expressing knowledge related to the forgetting task through self-generated texts. Furthermore, given the difficulty of identifying task-level correlations, LWF computes a data-wise *forgetting confidence* for each input by weighting the intended parameter updates with the Fisher Information Matrix. Based on this metric, LWF selects high-confidence data points and integrates machine unlearning techniques to periodically remove associated knowledge during the fine-tuning process.

To the best of our knowledge, LWF is the first systematic exploration to enhance the learning plasticity of pretrained generative language models by graceful forgetting. Through extensive experiments and analyses, we demonstrate the feasibility of improving fine-tuning performance through graceful forgetting. Hopefully, our empirical findings will contribute to a better understanding of this emerging topic and offer inspiration for future investigation and innovation.

2 Related Work

In this section, we provide an overview of how the critical concepts in our work are interpreted and applied across a broader research landscape, and analyze their commonalities and distinctions.

2.1 Negative Transfer

Negative transfer is a prevalent issue across multiple fields, with its interpretation and definition often varying depending on the context.

In Multi-Task Learning (MTL), negative transfer refers to the performance degradation caused by mutual interference among conflicting tasks (Go et al., 2023). While the primary objective of MTL is to learn multiple tasks simultaneously, methods aimed at mitigating negative transfer typically focus on quantifying inter-task relationships using sophisticated metrics like gradient directions (Jiang et al., 2023; Riemer et al., 2019) or signal-to-noise ratio (Go et al., 2023). Based on these metrics, tasks can be clustered into separate groups (Ramesh and Chaudhari, 2022) to reduce learning conflicts. In the MTL context, strategies that involve sacrificing part of the model's capacity, such as forgetting, are typically not regarded as appropriate solutions.

Another related domain is Continual Learning (CL), where different tasks are learned sequentially. In CL, negative transfer is bi-directional: it can refer to the interference of previously acquired knowledge with the learning of new tasks, or to the forgetting of past knowledge caused by learning new ones (Karakida and Akaho, 2022). While most CL methods focus on maintaining memory stability when learning new tasks (Kirkpatrick et al., 2017; Schwarz et al., 2018), recent studies have highlighted that this stability often comes at the cost of reduced learning plasticity. In response, these approaches actively weaken the preservation strength of past memory, seeking a balance between memory stability and learning plasticity (Liang and Li, 2023; Wang et al., 2021; Schwarz et al., 2018), which can be seen as an implicit form of forgetting.

In our context, the pretrain-finetune paradigm, negative transfer denotes the detrimental influence of certain pre-trained knowledge on the target fine-tuning task. A distinguishing characteristic of negative transfer in this paradigm is that the pre-training data is typically inaccessible during fine-tuning, which renders many countermeasures used in MTL and CL ineffective or inapplicable.

2.2 Graceful Forgetting

In many domains, forgetting is traditionally viewed as an undesirable phenomenon, reflecting a failure to retain previously acquired knowledge. This concern is particularly prominent in CL, where one of the primary objectives is to overcome catastrophic forgetting (Kirkpatrick et al., 2017).

However, 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). Drawing inspiration from neuroscience (Anderson and Hulbert, 2021), an increasing number of studies have explored the potential of improving learning plasticity through actively forgetting irrelevant or outdated knowledge. In the context of CL, Wang et al. (2021) proposed a synaptic expansion-convergence mechanism to selectively forget preserved knowledge. Liang and Li (2023) realized controllable learning plasticity through gradient projection. In the context of fine-tuning, Chen et al. (2019) and Shen et al. (2024) incorporate model structural shrinkage regulation to enable implicit forgetting. While most of these methods were initially designed for non-autoregressive tasks, Chen et al. (2023) extended graceful forgetting to the LLM pre-training stage, enhancing multi-lingual ability by refreshing models. Ni et al. (2024) proposed a forgetting-before-learning method to achieve knowledge edition for LLMs.

Despite these efforts, existing graceful forgetting approaches are either inapplicable or significantly less effective when adapted to promoting fine-tuning of generative language models.

2.3 Machine Unlearning

Machine unlearning has emerged as a vibrant and rapidly evolving research area focused on selectively removing specific data, patterns, or knowledge from trained models (Yao et al., 2023). In the context of generative language models, unlearning is frequently employed to align model behavior with human values—such as safeguarding user privacy (Patil et al., 2024), eradicating harmful or biased content (Liu et al., 2024), and mitigating hallucinations (Yao et al., 2023). A range of unlearning strategies have been proposed for this purpose, including gradient ascent (Jang et al., 2023), localization-informed unlearning (Jang et al., 2023), and influence function-based approaches (Jia et al., 2024), among others. While current unlearning research about generative language models 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, Learning With Forgetting (**LWF**). It 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. It is important to note, however, that the framework is task quantity-agnostic, which will be elaborated in 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. To address this issue, we exploit the inherent characteristics of generative language models as an 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 $\mathcal{D}_{self}$. This approach also enables LWF to utilize unlabeled datasets.

3.2 Evaluating Forgetting Confidence

Not all kinds of knowledge oblivion contribute positively to model adaptation. To ensure that forgetting improves, rather than degrades, the fine-tuning performance, we propose a confidence metric to evaluate the safety of unlearning specific knowledge. Considering the semantic richness of natural language, a task-level metric proves insufficiently nuanced (Pruksachatkun et al., 2020). Therefore, we define the *forgetting confidence* at the individual data point level, enabling fine-grained assessment of what 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

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

Here θ represents the model parameters. 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
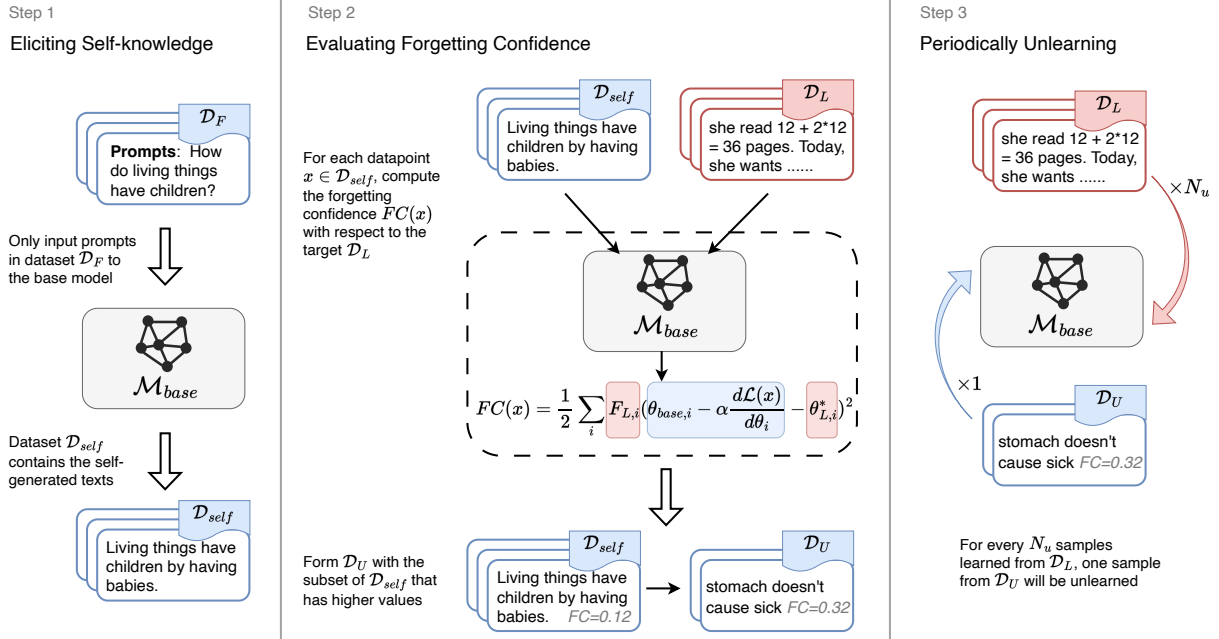
Figure 1: The overview of the LWF framework. Given the forgetting task $\mathcal{D}_F$ and learning task $\mathcal{D}_L$, LWF first constructs $\mathcal{D}_{self}$ through self-generated texts to represent the knowledge regarding the forgetting task. Then, with the Fisher Information Matrix F_L and the 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.

forgetting confidence as:

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

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 around θ_L^*:

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

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:

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

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}, and $\mathcal{L}(x)$

refers to the cross-entropy loss of x. α controls the margin of the single-step update. θ_L^* is obtained by training the base model on $\mathcal{D}_L$. We include a detailed step-by-step deduction in Appendix D and conduct a sensitivity analysis about the approximation error of the one-step update in Appendix F.

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

Due to the well-documented instability of machine unlearning (Yao et al., 2023; Liu et al., 2024), directly unlearning samples from $\mathcal{D}_{self}$ may yield inconsistent performance gains, particularly because the selected samples are only potentially conflicting with the target task, not definitively so. To mitigate the instability introduced by unlearning, we adopt a *"periodically unlearning"* strategy that interleaves learning and unlearning simultaneously throughout one training process.

Specifically, we introduce a fixed unlearning interval N_u, such that the unlearning is applied every

N_u learning steps. For instance, when $N_u = 7$, we first select a subset $\mathcal{D}_U$ from $\mathcal{D}_{self}$, consisting of the top $\frac{|\mathcal{D}_L|}{7}$ samples of the highest forgetting confidence scores $FC(x)$. During training, both of the learning dataset $\mathcal{D}_L$ and the unlearning dataset $\mathcal{D}_U$ are utilized: for every 7 learning samples drawn from $\mathcal{D}_L$, one sample from $\mathcal{D}_U$ is unlearned. This balanced interleaving helps prevent the forgetting process from impairing the learning dynamics.

We use *Gradient Ascent* (Bourtoule et al., 2021) as our unlearning algorithm, which merely involves negating the loss function. Specifically, for a periodic batch $\mathcal{X} = \{x_1^l, \ldots, x_{N_u}^l, x^u\}$ where $\{x_1^l, \ldots, x_{N_u}^l\} \subset \mathcal{D}_L$ and $x^u \in \mathcal{D}_U$, the loss can be written as:

$$\mathcal{L}_{pu}(\mathcal{X}) = \sum_{x \in \{x_1^l, \ldots, x_{N_u}^l\}} \mathcal{L}(x) - \beta \mathcal{L}(x^u) \quad (5)$$

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

4 Experiments

4.1 Setup

4.1.1 Datasets

To evaluate the effectiveness of LWF, we apply our method to domain-specific question-answering tasks. This choice is motivated by the well-established evaluation metrics (i.e., accuracy) and the relatively clear delineation of knowledge boundaries across domains. We further discuss task generalizability in Appendix C.

We select five datasets spanning diverse domains: **gsm8k** (Cobbe et al., 2021) for mathematical reasoning; **qasc** (Khot et al., 2020) for elementary science; **sst5** (Socher et al., 2013) for sentiment classification; **dental**, a subset of MedMCQA (Pal et al., 2022) focusing on dental knowledge; and **psychol**, a subset of MMLU (Hendrycks et al., 2021) covering psychology.

4.1.2 Baselines

As discussed in Sections 1 and 2, although research on forgetting in broader contexts is not lacking, few existing graceful forgetting methods are specifically designed to enhance fine-tuning plasticity for pre-trained generative language models, which makes effective method migration challenging. To place LWF in a broader methodological context, we select three baseline approaches originally developed for different settings but adaptable to ours with minimal modifications: BSS (Chen

	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*.

et al., 2019) and SRS (Shen et al., 2024), two structural regularization methods that mitigate negative transfer from pretrained knowledge in non-autoregressive models; and F-learning (Ni et al., 2024), an active unlearning technique proposed for model editing in large language models.

4.1.3 Implementation Details

We use Llama3.2-1B (Touvron et al., 2023) as the base model. For eliciting self-knowledge, we employ a 3-shot prompt concatenated with the input question, use greedy decoding, and limit the maximum number of generated tokens to 256. For computing the forgetting confidence, we set the one-step update coefficient $\alpha = 1 \times 10^{-2}$ (as defined in Equation 4). To maintain gradient coherence during periodic unlearning, we interleave the datasets $\mathcal{D}_L$ and $\mathcal{D}_U$ by including one sample from $\mathcal{D}_U$ for every N_u samples from $\mathcal{D}_L$. We set $N_u = 7$ and choose β (from Equation 5) as either 0.1 or 0.05, depending on the forgetting task. The training batch size is 4, which results in one unlearning sample appearing every two batches. We optimize using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-5}, and train for one full epoch. All experiments are conducted on eight NVIDIA RTX 4090 GPUs with full-parameter fine-tuning. For further implementation details, please refer to Appendix A and our source code repository https://github.com/rubickkcibur/LWF.

4.2 Results on Question Answering

Table 1 presents the results on question-answering tasks. Each column corresponds to a distinct learning task, and each row represents a specific forgetting scenario. For example, the value +5.38% at the intersection of the gsm8k column and the qasc row indicates that, when fine-tuning on gsm8k, unlearning qasc using LWF improves performance by

	gsm8k	qasc	sst5	dental	psychol
none	56.56	68.36	55.52	49.39	80.00
gsm8k	-	+5.37%	+0.97%	+0.78%	+2.06%
qasc	+1.61%	-	+0.74%	-3.06%	+0.46%
sst5	+7.37%	+2.68%	-	+0.00%	+2.75%
dental	+1.20%	+1.26%	+1.30%	-	+0.91%
psychol	+1.47%	+4.42%	+0.24%	-0.14%	-
mixed	+4.56%	+7.90%	+1.06%	+0.78%	+1.60%

Table 2: Results on domain-specific question answering with a larger language model.

	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
F-learning	17.29	46.44	51.31	34.07	42.75	38.37
LWF-mixed	21.08	45.36	50.59	37.41	50.10	40.91

Table 3: Accuracy results of three adapted baseline methods compared to vanilla fine-tuning and LWF under *mixed* setting.

5.38% compared to vanilla fine-tuning. In particular, the first row, labeled *none*, reports the vanilla fine-tuning results (*i.e.*, no unlearning). The last row, labeled *mixed*, represents a forgetting scenario in which all datasets except the target learning dataset are jointly unlearned. All entries except those in the *none* row report the percentage improvement in accuracy relative to the corresponding vanilla fine-tuning baseline.

As shown in the results, in most cases, LWF improves performance on the learning task compared to vanilla fine-tuning. Exceptions occur when learning dental while forgetting qasc or psychol, which we believe is attributed to the low forgetting confidence of self-generated samples. This is further evidenced by the results of the *mixed*, where consistent improvements across all learning tasks suggest that combining diverse forgetting datasets increases the pool of high-confidence candidates, thereby enhancing the likelihood of performance gains. *The mixed setting is also recommended in practice for stable effectiveness.*

Table 3 compares LWF-mixed with the three adapted baseline methods. As shown, although one of the baselines occasionally achieves the highest improvement in certain cases, on average, all three perform worse than both LWF-mixed and even vanilla fine-tuning. We attribute this to the fundamental mismatch between these methods and our setting. Specifically, BSS and SRS were designed

for non-autoregressive models and are not optimized for autoregressive generation tasks. Meanwhile, F-learning targets model editing scenarios in which the knowledge to be unlearned is explicitly known a priori to be detrimental or outdated, a condition that does not hold in our context, where forgetting targets are not necessarily harmful.

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. Details are discussed in Appendix B.

4.3 Scalability Analysis

A natural question is: *Would LWF become less effective in larger generative language models, given that their increased parameter capacity may better accommodate conflicting knowledge?* To investigate this, we apply LWF to the Llama3-8B model. As shown in Table 2, LWF can still improve fine-tuning performance in most cases. While the magnitude of relative improvement has declined overall, this trend is partly due to the stronger baseline performance of the larger model. Notably, the discipline gained from smaller model settings generalizes well to the larger: the *mixed* setting is still the best choice to achieve stable effectiveness.

4.4 Analysis on Forgetting Confidence

A central component of LWF is the computation of forgetting confidence, which estimates the likelihood that forgetting a specific data point will benefit the learning task. While we propose a computable metric $FC(x)$ (in Sec 3.2), it involves approximations and does not provide an exact mapping between gradient updates and performance outcomes. As a result, $FC(x)$ should be viewed as a heuristic rather than an absolute measure. In practice, we also observe that unlearning data with relatively low FC does not necessarily lead to performance degradation.

To gain deeper statistical insight into the role of forgetting confidence, we design an ablation study using an inverse filtering strategy: selecting data with the lowest forgetting confidence. For each learning task, we compute the percentage change in accuracy relative to vanilla fine-tuning, across all forgetting tasks. To increase the sample size, we aggregate results over four unlearning rates: $\beta \in \{0.05, 0.10, 0.20, 0.25\}$.

Fig 2 shows the distribution of accuracy changes under the two filtering strategies. The *red* region corresponds to unlearning data with the *highest*
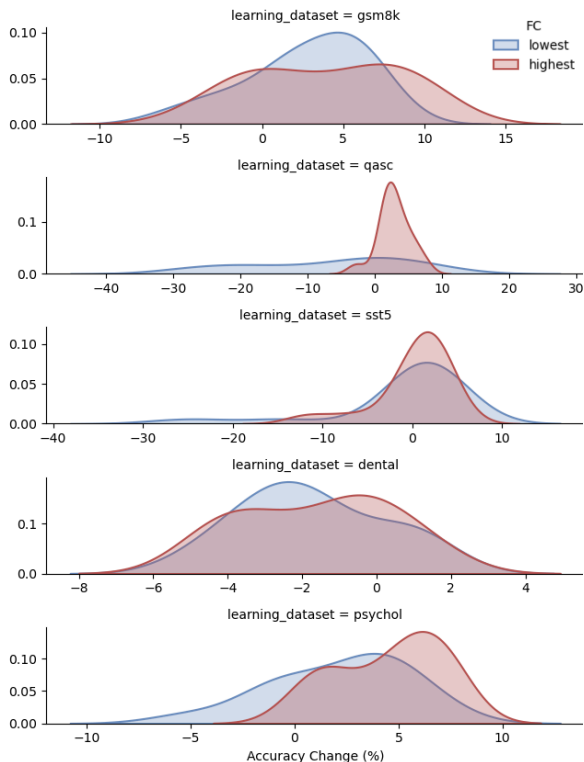
Figure 2: Distribution of accuracy changes between two filtering strategies. The *y*-axis represents distribution density. All percentages are calculated based on vanilla fine-tuning.

$\mathcal{D}_F$	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 4: Results of *ahead unlearning*, where unlearning is completed before fine-tuning, as an ablation study for periodically unlearning.

$\mathcal{D}_F$	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 5: Results of *randomly unlearning*, where unlearning is randomly executed during fine-tuning, as an ablation study for periodically unlearning.

FC, while the *blue* region represents unlearning data with the lowest. The *x*-axis indicates the percentage change in accuracy relative to vanilla fine-tuning. As shown, unlearning high-*FC* data generally outperforms the alternative in two aspects: higher average performance gain and greater stability. Specifically, high-*FC* unlearning yields smaller variance and a narrower overall range, whereas unlearning low-*FC* data leads to more volatile outcomes, including severe performance drops in extreme cases. In summary, prioritizing data with high forgetting confidence results in consistently better and more robust performance, making it a more reliable strategy in practice.

4.5 Ablation 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 effective strategy for integrating learning and unlearning.

We compare two alternative unlearning strategies. The first performs unlearning entirely before the learning process, which we refer to as *ahead*

unlearning. The second interleaves unlearning steps randomly during training, termed *randomly unlearning*. For fairness, all three strategies, including *periodically unlearning*, maintain the same ratio of learning to unlearning samples.

Table 4 and Table 5 present the results of *ahead unlearning* and *randomly unlearning* respectively. As we can see, both strategies are generally much less effective than *periodically unlearning* (Table 1), with most learning-forgetting combinations resulting in performance degradation. Notably, *ahead unlearning* exhibits several extremely detrimental cases, primarily attributed to the unintended disruption to foundational pre-trained knowledge caused by premature unlearning steps. If such critical knowledge is compromised, the subsequent fine-tuning process may suffer severe performance loss. In conclusion, the interleaved learning and unlearning combination is better than conducting them separately, and switching them periodically is superior to randomly.

4.6 Analysis on the Forgotten Task

In this section, we examine how model performance evolves on the forgetting task before and after applying LWF. While it may seem intuitive
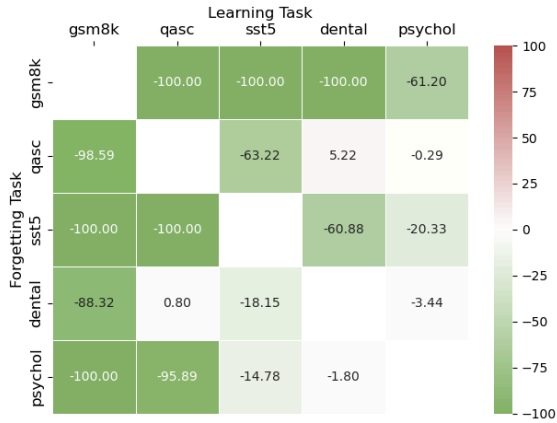
Figure 3: Accuracy change percentage of the forgetting task across different learning-forgetting combinations. Percentages are computed based on vanilla fine-tuning.
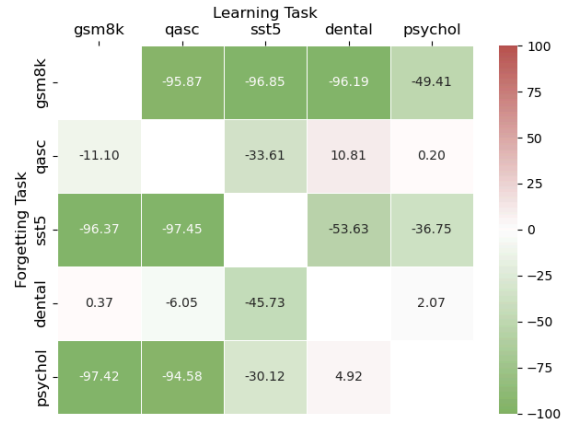

Figure 5: TTR change percentage of the forgetting task across different learning-forgetting combinations. Percentages are computed based on vanilla fine-tuning.
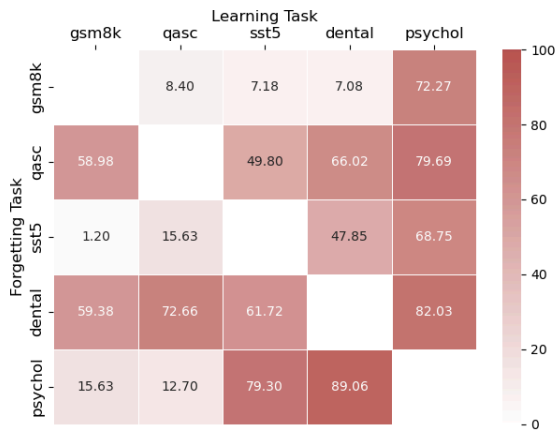

Figure 4: Cosine similarity between the outputs of forgetting tasks generated by the vanilla fine-tuned model and LWF resulting model. Values are multiplied by 100.

that the model's knowledge of the forgetting task would be substantially erased, the empirical results reveal a more nuanced picture. Figure 3 shows a heatmap of accuracy changes (in percentage points) on the forgetting tasks relative to vanilla fine-tuning. For instance, when the learning task is psychol and the forgetting task is gsm8k, the value −61.20 indicates a 61.2% drop in performance on gsm8k. As the figure illustrates, although accuracy generally declines across most learning-forgetting pairs, the extent of degradation varies significantly. In approximately half of the cases, performance drops by nearly 100%, suggesting near-complete unlearning; in others, the decline is much smaller, particularly when either the learning or the forgetting task involves dental or psychol.

Additionally, we noticed that the accuracy alone merely reflects whether the final answer is correct,

which is insufficient to fully capture the nuanced changes in the generated outputs of generative models. To gain deeper insights, we further analyze the semantic shifts in the responses. Specifically, we use SimCSE model (Gao et al., 2021) to obtain the sentence vectors of the responses generated by the vanilla fine-tuned model and LWF model when answering the same forgetting task questions. We then compute the cosine similarity between the resulting sentence embeddings to quantify the degree of semantic change. Results are presented in Fig 4.

Given that a cosine similarity score above 80% is typically required to confidently assert semantic similarity between two sentences, the results indicate that most learning-forgetting combinations exhibit substantial semantic changes. Notably, the cases where similarity approaches or exceeds this threshold largely coincide with those showing minimal accuracy drops, *i.e.*, combinations involving the dental or psychol. We believe this phenomenon may be attributed to that dental and psychol are inherently more complex than the other tasks. Acquiring or forgetting these domains likely requires engagement with richer, more structured knowledge, which in turn makes them more resistant to severe forgetting.

We also evaluate changes in lexical diversity, as shown in Fig 5, where Type Token Ratio (TTR) is the metric. Similar to the trends observed in accuracy changes, the TTR experiences a significant decline in most combinations, and the cases maintaining high accuracy and semantic similarity also largely preserve their lexical diversity.
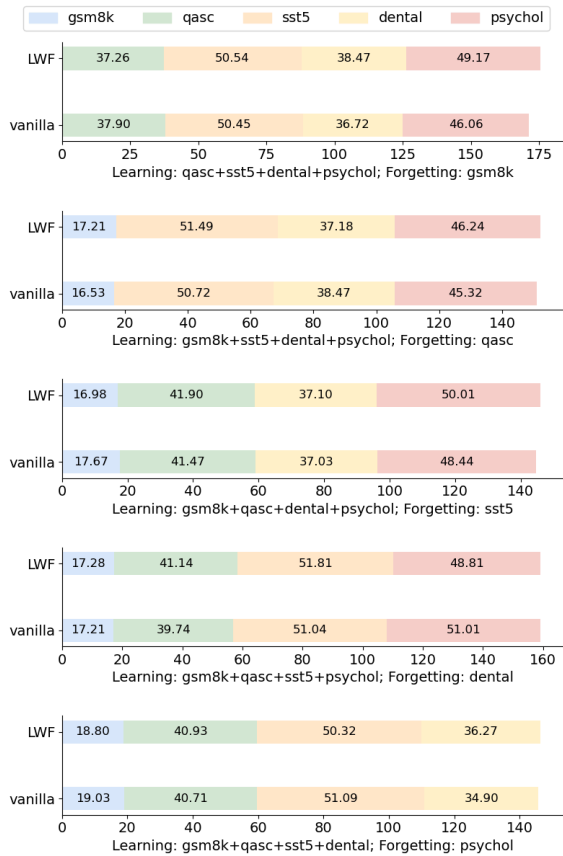
Figure 6: Accuracy results in the multi-task learning setting. Labeled below each subplot are the forgetting task and learning tasks.

4.7 Multi-Task Learning

In this section, we examine the performance of LWF in multi-task learning scenarios. Specifically, we treat one of the five datasets as the forgetting task and use the remaining four as learning tasks. To mitigate catastrophic forgetting of earlier tasks, we train on a balanced mixture of all learning datasets. Fig 6 compares the overall multi-task accuracy between LWF and vanilla fine-tuning. As we can see, while not all individual learning task gets improved, LWF outperforms vanilla fine-tuning in general.

These results highlight the complexity of multi-task learning. Although LWF enables overall performance gains through controlled forgetting, the interactions among learning tasks are non-trivial. Improvements in some tasks may come at the expense of others.

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 the 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

LWF still has several limitations that warrant further investigation. First, our proposed forgetting confidence metric is based on analyzing intended parameter updates to select data for unlearning. While empirical results demonstrate its statistical effectiveness, performance may degrade when the pool of candidate forgetting samples is small. Since quantifying interactions among training examples remains a longstanding challenge in knowledge transfer, we anticipate that future work will develop more precise and robust metrics for estimating forgetting confidence.

Additionally, computing forgetting confidence incurs non-negligible computational overhead, primarily due to the need to compute gradients for each candidate forgetting sample. As a result, the additional computational cost is approximately equivalent to one full training epoch on $\mathcal{D}_{self}$. However, since our method is designed for task-specific fine-tuning rather than large-scale pre-training, this overhead is generally acceptable in practical settings.

Acknowledgments

References

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

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *SP*, pages 141–159. IEEE.

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.

Tom Kocmi, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popovic. 2022. Findings of the 2022 conference on machine translation (WMT22). In *WMT*, pages 1–45. Association for Computational Linguistics.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2024. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. In *ACL (1)*, pages 5716–5731. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

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.

Rahul Ramesh and Pratik Chaudhari. 2022. Model zoo: A growing brain that learns continually. In *ICLR*. OpenReview.net.

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.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

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

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2023b. 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.

A   More Implementation Details

Tables 6, 7, 8, 9, and 10 present the few-shot Chain-of-Thought (CoT) prompts designed for each dataset, which are used during both self-knowledge elicitation and evaluation. As shown in the prompts, answers are formatted with the phrase *"The answer is"* to facilitate automated answer extraction. Any model output that does not conform to this format is considered incorrect. When multiple instances of *"The answer is"* appear in the output, the first occurrence is taken as the final answer.

B   Side Effect

Although in Sec 4 we have verified that LWF can improve the target fine-tuning task by sacrificing the unlearning task, it is unknown how this procedure will influence unintended tasks that are neither part of the learning task nor the forgetting task (for simplicity, we denote them as side-tasks). In this section, we discuss the side effects of LWF.

Firstly, we notice a *superficial forgetting* problem, which happens between gsm8k and the other four datasets. Specifically, while all five datasets used in our experiments are question-answering datasets, gsm8k is a free-form numerical QA dataset, but the other four are multiple-choice QA datasets (see examples in Table 6 and Table 7). We observed that this format discrepancy can lead to significant performance degradation on side-tasks 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. An illustration is provided in Table 11. As it shows, although the rationale portion of the output appears coherent, the resulting model fails to select a valid option at the end of its response.

This *superficial forgetting* suggests that, when applying LWF, 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 in practice.

Apart from the superficial forgetting issue introduced by gsm8k, we also compute the average accuracy of side-tasks on combinations of the other four datasets, and compare it to that of the vanilla fine-tuned model. Table 12 shows the results, where each column represents the learning task and the row indicates the forgetting task. As observed, the side effects vary depending on the specific learning-forgetting combinations. In general, the impact is much milder when learning complex tasks like psychol and dental.

C   Task Generalizability

In Sec. 4, we evaluate the effectiveness of LWF on the domain-specific question-answering task. This is primarily motivated by its well-established evaluation metrics and delineated knowledge boundaries, which help isolate the impact of graceful forgetting by minimizing confounding variables. Additionally, to examine the task generalizability, we conducted experiments in two other settings, both related to the multi-lingual capability of generative language models.

First, we apply LWF to machine translation tasks, We selected four language datasets–German (DE), French (FR), Turkish (TR), and Chinese (ZH)–from the WMT (Kocmi et al., 2022) corpus to evaluate the model's ability to translate from English into those languages. BLEU (Papineni et al., 2002) and BERTScore-F1 (Zhang et al., 2020) are used as evaluation metrics. As Table 13 shows, while LWF continues to yield performance gains in most cases, the extent of these improvements is considerably less pronounced compared to QA tasks. In addition, the results reflected by the two metrics are not entirely consistent, which is mainly due to the incompleteness of evaluation metrics in machine translation tasks.

The second experiment was conducted on the multi-lingual QA task. Specifically, we applied LWF to the gsm8k dataset across five different languages: English (EN), Italian (IT), Chinese (ZH), Spanish (ES), and Turkish (TR). From the results presented in Table 14, 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, and 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 helps stabilize LWF's performance when English is involved.

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 6: 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 7: 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 8: 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 9: 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 10: 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 11: Cases of superficial forgetting.

	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 12: Average accuracy changes on side-tasks after applying LWF. Percentages are calculated relative to vanilla fine-tuning.

	DE		FR		TR		ZH	
	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1
none	23.02	83.03	24.82	84.01	14.35	66.99	16.67	79.04
DE	-	-	+0.12%	-0.05%	-0.42%	+0.06%	+0.00%	+0.04%
FR	+0.48%	-0.02%	-	-	+0.21%	+0.24%	+0.06%	+0.05%
TR	+0.26%	+0.04%	+0.00%	+0.01%	-	-	+0.18%	+0.09%
ZH	+0.13%	+0.01%	+0.40%	+0.00%	-0.07%	+0.34%	-	-

Table 13: Results on machine translation. Each column shares the same learning target language, and the rows represent different forgetting languages. All percentages are calculated relative to *none*

	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 14: Results on multi-lingual question-answering. All percentages are calculated relative to *none*.

	gsm8k	qasc	sst5	dental	psychol
vanilla-FT	60.96	66.06	54.71	43.17	78.72
LWF-mixed	+1.23%	+0.17%	+1.57%	+0.19%	+1.16%

Table 15: Accuracy results of LWF-mixed with Qwen2.5-1.5B as the base model.

These two experiments inspire us that graceful forgetting mechanisms may not transfer homogeneously across all NLP tasks. Effective performance on complicated tasks may require task-specific adaptations and rigorous per-task analysis.

D Deduction of Forgetting Confidence

The second-order Taylor expansion of $\log P(\theta|\mathcal{D}_L)$ around θ_L^* is:

$$\log P(\theta|\mathcal{D}_L) =$$
$$\log P(\theta_L^*|\mathcal{D}_L) + \left(\frac{\partial \log P(\theta|\mathcal{D}_L)}{\partial \theta}\Big|_{\theta_L^*}\right)(\theta - \theta_L^*)$$
$$+ \frac{1}{2}(\theta - \theta_L^*)^T \left(\frac{\partial^2 \log P(\theta|\mathcal{D}_L)}{\partial^2 \theta}\Big|_{\theta_L^*}\right)(\theta - \theta_L^*)$$
$$+ R_2(\theta)$$

(6)

where $R_2(\theta)$ is the higher-order term and is neglected. Note that we define $\theta_L^* = \arg\max P(\theta|\mathcal{D}_L)$, which implies $\left(\frac{\partial \log P(\theta|\mathcal{D}_L)}{\partial \theta}\Big|_{\theta_L^*}\right) = 0$. Therefore, the first-order (linear) term in the Taylor expansion vanishes. Furthermore, since forgetting confidence is used only to rank data samples, and the constant term does not affect the ranking results, we omit it as well. As a result, Eq. 3 retains only the second-order term. Following prior work (Kirkpatrick et al., 2017), we replace the Hessian matrix in Eq. 3 with the Fisher information matrix, which can be interpreted as the negative expected value of the Hessian under the model distribution:

$$F_L = \mathbb{E}\left[\left(-\frac{\partial^2 \log P(\theta|\mathcal{D}_L)}{\partial^2 \theta}\right)\Big|_{\theta_L^*}\right] \quad (7)$$

To efficiently measure the resulting influence of sample x, we use a single-step update from the base model to represent $\theta^*(x)$:

$$\theta^*(x) \approx \theta_{base} - \alpha \frac{\partial \mathcal{L}(x)}{\partial \theta} \quad (8)$$

steps	1	2	3	4
overlapping	100.00%	99.81%	99.91%	99.91%

Table 16: $\mathcal{D}_U$ overlapping with different approximation steps (Equation 4)

By substituting Eq. 7 and Eq. 8 into Eq. 3, we can get Eq. 4

E Architecture Transferability

To evaluate the architectural transferability of LWF, we assess its performance on the Qwen2.5-1.5B model. As shown in Table 15, LWF-mixed consistently outperforms vanilla fine-tuning. However, compared to the results obtained with Llama models, the performance gain is relatively smaller—likely due to the stronger intrinsic capabilities of the Qwen model, which may reduce the relative benefit of plasticity enhancement.

F Approximation Error Study

In Equation 4, we employ a one-step update to approximate the optimal model parameter $\theta^*(x)$. Since the forgetting confidence is highly sensitive to the accuracy of this approximation, we conduct an empirical study to quantify the estimation error introduced by this simplification. Specifically, we compare the one-step update with multi-step updates (2, 3, and 4 steps), evaluating the similarity of the resulting unlearning dataset $\mathcal{D}_U$ by computing the proportion of overlapping elements. As shown in Table 16, the results indicate that multi-step updates yield only marginal improvements over the one-step approximation, suggesting that the simpler approach is sufficiently accurate in practice.