LLM Surgery: Efficient Knowledge Unlearning and Editing in Large Language Models

Anonymous ACL submission

Abstract

LLMs have revolutionized various domains, yet their utility comes with significant challenges related to outdated or problematic knowledge embedded during pretraining. This paper addresses the challenge of modifying LLMs to unlearn problematic and outdated information while efficiently integrating new knowledge without retraining from scratch. We propose LLM Surgery, a framework to efficiently modify LLM behaviour by optimizing a three component objective function that: (1) Performs reverse gradient on unlearning dataset (problematic and outdated information), (2) Performs gradient descent on the update dataset (new and updated information), and (3) Minimizes the KL divergence on the retain dataset (small subset of unchanged text), ensuring alignment between pretrained and modified model outputs. Due to the lack of publicly available datasets for our novel task, we introduce a new corpora and an evaluation benchmark. Using Llama2-7B, we demonstrate that LLM Surgery can achieve significant forgetting on the unlearn set, a 20% increase in accuracy on the update set, and maintain performance on the retain set. Code and dataset are available here.

1 Introduction

LLMs excel across various tasks (Vilar et al., 2022; Zhu et al., 2023; Zhang et al., 2023b), but they rely on extensive internet-sourced data (Xu et al., 2024; Achiam et al., 2023a), which includes problematic content such as copyrighted material and personal information (Zhao et al., 2022). This can lead to LLMs inadvertently memorizing and generating such data (Smith et al., 2023). Additionally, LLMs are limited by the temporal scope of their training data, lacking awareness of events or updates after training (Lazaridou et al., 2021; Agarwal and Nenkova, 2022), which risks generating outdated or incorrect information. Retraining LLMs to address these issues is expensive; for example, pretraining Llama-3 requires 7.7 million GPU hours and produces significant CO_2 emissions (AI@Meta, 2024).

To ensure LLMs remain relevant and beneficial, it is essential to develop methods that allow models to **(1)** unlearn problematic or outdated data, **(2)** incorporate new information, and **(3)** maintain performance on standard benchmarks. Else, LLMs pose legal and ethical risks (Zhang et al., 2023a), as shown by lawsuits filed against companies using copyrighted material for training (Times, 2023). As LLMs become more widespread, the risk of perpetuating outdated or legally sensitive content grows, leading to severe consequences (Zhao et al., 2022).

To address these challenges, we propose 'LLM Surgery' for efficiently updating LLM knowledge to ensure it remains current, accurate, and legally compliant. Our contributions are: **(1)** the development of an objective function that applies reverse gradient to unlearn data, gradient descent to new data, and minimizes KL divergence on retained data, ensuring alignment between pretrained and updated outputs. Unlike prior works, this objective function can operate in continual-pretraining regime, eliminating the need for expensive annotations of input-output pairs, thereby enabling efficient and scalable unlearning and updating of LLM knowledge; **(2)** the creation of a novel and comprehensive dataset with unlearn, update, and retain sets for both training and evaluation; and **(3)** analysis showing significant forgetting on unlearn set, a 20% accuracy improvement on update set, and stable performance on the retain set.

2 Proposed Methodology

Given the unlearn dataset ($\mathcal{D}^{\text{unl}}$), the update dataset ($\mathcal{D}^{\text{upd}}$), and the retain dataset ($\mathcal{D}^{\text{rtn}}$), our goal is to modify a LLM, $\boldsymbol{\theta}^{\text{pre}}$, so that the resulting LLM, $\boldsymbol{\theta}^{\text{surgery}}$, meets the following criteria: (1) it unlearns specific information from $\boldsymbol{\theta}^{\text{pre}}$ based on $\mathcal{D}^{\text{unl}}$, (2) updates $\boldsymbol{\theta}^{\text{pre}}$ with new knowledge from $\mathcal{D}^{\text{upd}}$, and (3) maintains the same level of performance as $\boldsymbol{\theta}^{\text{pre}}$ on standard benchmark tasks using $\mathcal{D}^{\text{rtn}}$. To achieve these goals, we propose the LLM Surgery.

LLM Surgery Objective optimizes a objective function, as defined in Equation 1. This function is designed to accomplish three key objectives: (1) ***Unlearning***: The first term in Equation 1 encourages the model to unlearn specific information by performing gradient ascent on $\mathcal{D}^{\text{unl}}$. This ensures that the model parameters are optimized along the reverse gradient direction of the learned patterns in $\mathcal{D}^{\text{unl}}$, effectively erasing targeted knowledge, (2) ***Updating***: The second term facilitates learning of new information by applying gradient descent on the $\mathcal{D}^{\text{upd}}$. This ensures the model integrates the updated knowledge, and (3) ***Retention***: The third term minimizes the KL-divergence between the output distributions of $\boldsymbol{\theta}^{\text{surgery}}$ and $\boldsymbol{\theta}^{\text{pre}}$ on $\mathcal{D}^{\text{rtn}}$. This ensures that the model retains essential knowledge that should remain unaffected by LLM Surgery. Mathematically, it can be expressed as follows:

$$
\begin{aligned}
\mathcal{L} = & \underbrace{\frac{\lambda_1}{|\mathcal{D}^{\text{unl}}|} \sum_{i=1}^{|\mathcal{D}^{\text{unl}}|} \frac{1}{|x_i^{\text{unl}}|} \sum_{t=2}^{|x_i^{\text{unl}}|} l\Big(P[\cdot \mid x_{i;<t}^{\text{unl}}; \boldsymbol{\theta}^{\text{surgery}}], \mathbf{e}_t^{\text{unl}}\Big)}_{\texttt{unlearn}} \\
& - \underbrace{\frac{\lambda_2}{|\mathcal{D}^{\text{upd}}|} \sum_{i=1}^{|\mathcal{D}^{\text{upd}}|} \frac{1}{|x_i^{\text{upd}}|} \sum_{t=2}^{|x_i^{\text{upd}}|} l\Big(P[\cdot \mid x_{i;<t}^{\text{upd}}; \boldsymbol{\theta}^{\text{surgery}}], \mathbf{e}_t^{\text{upd}}\Big)}_{\texttt{update}} \\
& - \underbrace{\frac{\lambda_3}{|\mathcal{D}^{\text{rtn}}|} \sum_{i=1}^{|\mathcal{D}^{\text{rtn}}|} \frac{1}{|x_i^{\text{rtn}}|} \sum_{t=2}^{|x_i^{\text{rtn}}|} \text{KL}\Big(P[\cdot \mid x_{i;<t}^{\text{rtn}}; \boldsymbol{\theta}^{\text{surgery}}] \Big\| P[\cdot \mid x_{i;<t}^{\text{rtn}}; \boldsymbol{\theta}^{\text{pre}}]\Big)}_{\texttt{retain}}
\end{aligned}
$$

where, $1 \leq i \leq |\mathcal{D}|$ is the ith example, and $|\mathcal{D}|$ is the dataset size. $x_i^{\text{unl}}(\in \mathcal{D}^{\text{unl}})$, $x_i^{\text{upd}}(\in \mathcal{D}^{\text{upd}})$, and $x_i^{\text{rtn}}(\in \mathcal{D}^{\text{rtn}})$ represent the sequences of tokens. λ_1, λ_2, and λ_3 are hyperparameters. $P[\cdot \mid x_{i;<t}; \boldsymbol{\theta}]$ is the predicted probability distribution over the possible next tokens at position t, given preceding tokens $x_{<t}$ and model parameters $\boldsymbol{\theta}$. $\mathbf{e}_t$ is the one-hot encoded representation of the true token at position t, and $l(\cdot, \cdot)$ is cross-entropy loss.

3 Experimental Setup

We conduct experiments using Llama2-7B (Touvron et al., 2023) to evaluate the effectiveness of the LLM Surgery in unlearning information from $\mathcal{D}^{\text{unl}}$, updating the model with new knowledge from $\mathcal{D}^{\text{upd}}$, and retaining performance on $\mathcal{D}^{\text{rtn}}$. As a preliminary step, we continually-pretrain *off-the-shelf* Llama2-7B on both $\mathcal{D}^{\text{unl}}$ and $\mathcal{D}^{\text{rtn}}$. The purpose of this continual-pretraining phase is to ensure that the model possesses the knowledge of $\mathcal{D}^{\text{unl}}$ and $\mathcal{D}^{\text{rtn}}$ prior to performing LLM Surgery. We refer to this continually-pretrained model as $\boldsymbol{\theta}^{\text{pre}}$.

Baseline Model: A naive approach would involve continually-pretraining *off-the-shelf* Llama2-7B using only the update dataset $\mathcal{D}^{\text{upd}}$ and the *full* retain dataset $\mathcal{D}^{\text{rtn}}$. This model is referred to as $\boldsymbol{\theta}^{\text{baseline}}$. However, this approach is both computationally expensive and time-intensive, particularly because $\mathcal{D}^{\text{rtn}}$ is significantly larger than $\mathcal{D}^{\text{upd}}$. The challenge is to achieve desired modifications efficiently without loss in overall model performance.

LLM Surgery: To overcome these challenges more efficiently, LLM Surgery leverages the general-purpose knowledge from $\mathcal{D}^{\text{rtn}}$ already embedded in $\boldsymbol{\theta}^{\text{pre}}$ to make targeted modifications, resulting in the modified model $\boldsymbol{\theta}^{\text{surgery}}$. Instead of starting from the *off-the-shelf* Llama2-7B model (as done when training $\boldsymbol{\theta}^{\text{baseline}}$), LLM Surgery begins with $\boldsymbol{\theta}^{\text{pre}}$. This approach efficiently unlearns specific information using $\mathcal{D}^{\text{unl}}$, integrates new knowledge using $\mathcal{D}^{\text{upd}}$, and preserves performance on critical tasks using a *small* subset of $\mathcal{D}^{\text{rtn}}$. Crucially, it avoids the high computational costs of full retraining from scratch (or in our context, retraining the Llama2-7B model on both $\mathcal{D}^{\text{upd}}$ and the *entire* $\mathcal{D}^{\text{rtn}}$ datasets as in the baseline model).

4 Dataset

Given the lack of publicly available datasets tailored for our task, we introduce a novel corpora and evaluation benchmark. This evaluation benchmark includes multiple-choice question-answer (MCQA) examples generated using GPT-4[1], which consist of questions, three answer choices, and labels indicating the correct choice. Accuracy is used as the evaluation metric. Below, we describe the construction of the corpora and characteristics of the evaluation benchmark (additional details in Appendix A):

Unlearn Dataset ($\mathcal{D}^{\text{unl}}$): We derive the unlearning corpora from the original Wikipedia biographies in the WikiBio GPT-3 Hallucination Dataset (Manakul et al., 2023), consisting of 238 *true* biographies. This dataset is divided into two subsets: (1) a subset of 119 biographies (50k tokens) used to simulate problematic data, such as copyrighted material, and (2) a subset of 119 biographies (50k tokens) representing outdated information. To assess the effectiveness of the LLM Surgery to unlearn this content from $\boldsymbol{\theta}^{\text{pre}}$, we generate 2,400 MCQA examples based on the unlearn corpora and use them for evaluating $\boldsymbol{\theta}^{\text{surgery}}$'s performance on the unlearning set.

Update Dataset ($\mathcal{D}^{\text{upd}}$): The update corpora is constructed by generating 119 fictitious biographies (80k tokens) in Wikipedia style using GPT-

[1]The prompt is inspired from Maini et al. (2024). 40:60 split is used for validation and test sets.

4 (Achiam et al., 2023b). These biographies correspond to the same 119 subjects in the outdated subset of the unlearning corpora. By generating these fictitious biographies, we ensure that $\boldsymbol{\theta}^{\text{pre}}$ has no prior exposure to this content before undergoing LLM Surgery. To assess the effectiveness of the LLM Surgery framework in incorporating this new information, we generate 2,400 MCQA examples based on the update set and used them for evaluating $\boldsymbol{\theta}^{\text{surgery}}$'s performance.

Retain Dataset ($\mathcal{D}^{\text{rtn}}$): To ensure that the modifications to the model do not degrade the performance on unrelated tasks, we create a retain corpora comprising 2 million tokens from OpenWebText dataset (Gokaslan et al., 2019) and 1 billion tokens from RedPajama-v1 dataset (Computer, 2023). While both $\boldsymbol{\theta}^{\text{pre}}$ and $\boldsymbol{\theta}^{\text{baseline}}$ utilize the *entire* retain corpora, $\boldsymbol{\theta}^{\text{surgery}}$ only uses a *small* subset of $\mathcal{D}^{\text{rtn}}$ to preserve the model's original performance on unrelated tasks. To assess the effectiveness of LLM Surgery in retaining performance, we developed an evaluation benchmark consisting of 20,000 MCQA examples from the 2 million tokens in the OpenWebText dataset to ensure that $\boldsymbol{\theta}^{\text{surgery}}$'s performance remains consistent with $\boldsymbol{\theta}^{\text{pre}}$.

5 Results

Table 1: Comparison of accuracy between $\boldsymbol{\theta}^{\text{pre}}$ and $\boldsymbol{\theta}^{\text{surgery}}$ across the unlearn, update, and retain datasets.

Method	Unlearn	Update	Retain
Llama2-7B	45.81	46.55	46.78
$\boldsymbol{\theta}^{\text{pre}}$	60.86	44.57	63.02
$\boldsymbol{\theta}^{\text{surgery}}$	41.86	63.32	60.4
Ablation Studies			
$\boldsymbol{\theta}^{\text{gd}}$	51.23	66.78	55.31
$\boldsymbol{\theta}^{\text{gd+kl}}$	62.26	61.35	63.15

Effectiveness of LLM Surgery In Table 1, we empirically demonstrate the effectiveness of the LLM Surgery framework by analyzing the accuracy of the continually-pretrained Llama2-7B model across the unlearn, update, and retain evaluation benchmarks, both before ($\boldsymbol{\theta}^{\text{pre}}$) and after ($\boldsymbol{\theta}^{\text{surgery}}$) applying LLM Surgery. The results indicate that $\boldsymbol{\theta}^{\text{surgery}}$ shows a 19% reduction in accuracy compared to $\boldsymbol{\theta}^{\text{pre}}$ on the unlearn benchmark. Additionally, on the unlearn benchmark, $\boldsymbol{\theta}^{\text{surgery}}$ exhibits 4% lower accuracy compared to the *off-the-shelf* Llama2-7B model, further confirming that $\boldsymbol{\theta}^{\text{surgery}}$ has successfully unlearned the targeted information. Simultaneously, we observe that $\boldsymbol{\theta}^{\text{surgery}}$ shows a 20% improvement in accuracy on the update benchmark compared to $\boldsymbol{\theta}^{\text{pre}}$, demonstrating LLM Surgery's ability to incorporate new knowledge. Crucially, $\boldsymbol{\theta}^{\text{surgery}}$ maintains comparable performance on the retain benchmark relative to $\boldsymbol{\theta}^{\text{pre}}$, suggesting that the LLM Surgery does not degrade the model's knowledge of unrelated tasks.

Efficiency of LLM Surgery To assess the efficiency gains of LLM Surgery, we compare its runtime to the naive approach used to obtain $\boldsymbol{\theta}^{\text{baseline}}$. Empirically, we observe that $\boldsymbol{\theta}^{\text{surgery}}$ achieves the same level of performance as $\boldsymbol{\theta}^{\text{baseline}}$, but with a 35x reduction in GPU hours. This significant speed-up is due to the fact that, in practice, the unlearn and update corpora are much smaller than the retain corpora. Unlike the naive approach of retraining from scratch ($\boldsymbol{\theta}^{\text{baseline}}$) using the *entire* retain corpora, LLM Surgery uses only 2% of the retain corpora, aided by the KL divergence term in the proposed objective function. These efficiency gains become even more pronounced as the size of the retain corpora increases, highlighting the scalability and effectiveness of LLM Surgery for larger datasets. Next, we evaluate the contribution of each term in LLM Surgery objective function.

Ablation Study 1: LLM Surgery with Gradient Descent Only ($\boldsymbol{\theta}^{\text{gd}}$) We remove the gradient ascent and KL-divergence terms, applying only the gradient descent term. As expected, relying solely on gradient descent for the update corpora improved accuracy of $\boldsymbol{\theta}^{\text{gd}}$ on the update benchmark from 44.6% to 66.8%, compared to $\boldsymbol{\theta}^{\text{pre}}$ (shown in Table 1). However, this approach resulted in $\boldsymbol{\theta}^{\text{gd}}$ having a 10% higher accuracy on the unlearn benchmark compared to $\boldsymbol{\theta}^{\text{surgery}}$, indicating insufficient unlearning. Moreover, accuracy on the retain set dropped by 8% relative to $\boldsymbol{\theta}^{\text{pre}}$, suggesting that using gradient descent alone can lead to undesirable forgetting of previously learned tasks.

Ablation Study 2: LLM Surgery with Gradient Descent and KL-Divergence ($\boldsymbol{\theta}^{\text{gd+kl}}$) To address the forgetting issue on $\mathcal{D}^{\text{rtn}}$, in this ablation, we include both the gradient descent and KL-divergence terms, omitting the gradient ascent term. This approach preserves the performance of $\boldsymbol{\theta}^{\text{gd+kl}}$ on $\mathcal{D}^{\text{rtn}}$, compared to $\boldsymbol{\theta}^{\text{pre}}$. However, it fails to unlearn the targeted information, as shown by the lack of accuracy drop on the unlearn benchmark. The influence of the KL-divergence term likely prevents effective unlearning. This outcome is undesirable, as $\boldsymbol{\theta}^{\text{gd+kl}}$ retains the information it was meant to unlearn. In other words, the gradient ascent term

is essential for effective unlearning, as it enables $\boldsymbol{\theta}^{\text{surgery}}$ to achieve a 19% reduction in accuracy on the unlearn benchmark, compared to $\boldsymbol{\theta}^{\text{pre}}$, proving that all three terms are necessary for efficiently and effectively modifying LLM behavior.

Activation Pattern Analysis: As shown in Appendix Figure 1, the activation patterns remain consistent across the three corpora—$\mathcal{D}^{\text{unl}}$, $\mathcal{D}^{\text{upd}}$, and $\mathcal{D}^{\text{rtn}}$—for both the *off-the-shelf* Llama2-7B and $\boldsymbol{\theta}^{\text{pre}}$. This is likely because Llama2-7B was pretrained on original Wikipedia biographies ($\mathcal{D}^{\text{unl}}$), as well as OpenWebText and RedPajama-v1 ($\mathcal{D}^{\text{rtn}}$). The same knowledge is reinforced in $\boldsymbol{\theta}^{\text{pre}}$ prior to LLM Surgery, resulting in highly similar activation patterns for *off-the-shelf* Llama2-7B and $\boldsymbol{\theta}^{\text{pre}}$. After LLM Surgery, however, $\boldsymbol{\theta}^{\text{surgery}}$ shows distinct activation patterns across $\mathcal{D}^{\text{unl}}$ and $\mathcal{D}^{\text{upd}}$, as LLM Surgery modifies $\boldsymbol{\theta}^{\text{pre}}$ to unlearn $\mathcal{D}^{\text{unl}}$ and incorporate $\mathcal{D}^{\text{upd}}$, while preserving its performance on $\mathcal{D}^{\text{rtn}}$. As a result, the activation patterns on $\mathcal{D}^{\text{rtn}}$ remain similar between $\boldsymbol{\theta}^{\text{pre}}$ and $\boldsymbol{\theta}^{\text{surgery}}$.

6 Related Works

Limitations of Existing Methods: Various methods have been proposed for LLM unlearning (Liu et al., 2024; Blanco-Justicia et al., 2024; Eldan and Russinovich, 2023; Chen and Yang, 2023; Maini et al., 2024; Thaker et al., 2024), each addressing the challenge of removing harmful or inaccurate content. One notable approach (Yao et al., 2023a) shares our goal of unlearning harmful or hallucinated information but heavily relies on fine-tuning the model with costly input-output annotations. Moreover, it requires an extensively annotated dataset to ensure coherence after unlearning, limiting the scalability of the method. Another line of research focuses on editing *structured* knowledge (Deng et al., 2024) by modifying model weights (Wei et al., 2024; Mitchell et al., 2021; Rozner et al., 2024), adjusting model architectures (Hartvigsen et al., 2024), or localizing knowledge within specific layers (Dai et al., 2021; Meng et al., 2022a,b). However, these approaches are typically effective only for a limited set of concepts (Li et al., 2023; Yao et al., 2023b) and fail to address the unlearning of outdated information comprehensively, leaving models susceptible to information extraction attacks (Patil et al., 2023).

Key Contributions of the Proposed Method: Our approach takes a significant step forward by unifying unlearning and model editing in a single framework, allowing for the removal of problematic data while simultaneously updating the model with new information. Unlike prior methods, our approach eliminates the need for costly annotations by leveraging a continual-pretraining objective, which makes it scalable to unstructured data and applicable across large token sets and concepts. Furthermore, our method addresses legal requirements such as the Right to be Forgotten Act (Zhang et al., 2023a), which demands the erasure of specific knowledge from LLM weights, something that fine-tuning or in-context learning approaches cannot effectively achieve.

Limitations of Existing Benchmarks and Our Dataset Proposal: The literature also suffers from a lack of publicly available datasets that can accurately assess the effectiveness of LLM unlearning and editing techniques. For example, Hartvigsen et al. (2024) attempts to reduce hallucinations by editing a pretrained LLM to produce *accurate* Wikipedia biographies. However, this evaluation is flawed, as the model likely encountered these *accuracte* biographies during pretraining, which introduces bias and makes it difficult to assess the effectiveness of the editing process. To address these shortcomings, we introduce a new corpora and evaluation set generated using GPT-4, where the updated information was unseen by the model before editing. Additionally, current benchmarks like TOFU (Maini et al., 2024) and WMDP (Li et al., 2024) are limited to short question-answer tasks, focusing only on unlearning. Our benchmark, by contrast, supports a more comprehensive evaluation of unified unlearning and model updating approaches in the continual-pretraining regime, facilitating more accurate assessments of these methods.

7 Conclusion

We propose LLM Surgery, an efficient framework for unstructured knowledge unlearning and editing in LLMs without retraining from scratch. LLM Surgery: **(1)** Performs reverse gradient on the unlearning set, **(2)** Performs gradient descent on the update set, and **(3)** Minimizes the KL divergence on the retain set, ensuring alignment between the pretrained and modified model outputs. Due to the lack of publicly available datasets tailored for our novel task, we compile a new corpora and an evaluation benchmark. On Llama2-7B, LLM Surgery can achieve significant forgetting on unlearn set, a 20% increase in accuracy on update set, and maintain performance on retain set.

8 Limitations

While the proposed LLM Surgery framework shows promise in addressing the challenges of unlearning outdated information and integrating new knowledge into LLMs, the paper has several limitations:

- **Limited Benchmarking Datasets:** While the paper introduces a new corpora and evaluation benchmark, the dataset is limited to the specific task of assessing LLM Surgery's performance in unlearning and updating Wikipedia-style biographies. The generalizability of the LLM Surgery framework's results to other types of knowledge remains uncertain. The lack of broader, publicly available datasets hampers our ability to comprehensively evaluate LLM Surgery across diverse domains. With our proposed dataset and evaluation benchmark, we aim to help address this gap in the literature.

- **No Exploration of Catastrophic Forgetting:** While LLM Surgery aims to maintain performance on the retain set, the long-term effects of iterative unlearning and updating on overall model stability are not fully explored. The paper focuses on short-term performance retention, but catastrophic forgetting—where updates gradually degrade previously learned knowledge over multiple iterations—remains an open concern.

- **Ethical and Practical Implications:** The paper does not fully explore the ethical implications of unlearning certain pieces of knowledge. The decision of what constitutes "problematic" information is inherently subjective and could vary across different contexts or stakeholders. This raises concerns about the potential misuse of LLM Surgery for selectively modifying a model's behavior in ways that could reinforce bias or misinformation if not properly governed. Additionally, the paper does not address the potential risk of unintended forgetting—where critical but unrelated knowledge might be inadvertently lost during the unlearning process.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023b. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

AI@Meta. 2024. Llama 3 model card.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2024. Digital forgetting in large language models: A survey of unlearning methods. *arXiv preprint arXiv:2404.02062*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset. https://github.com/togethercomputer/RedPajama-Data.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Unke: Unstructured knowledge editing in large language models. *arXiv preprint arXiv:2405.15349*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2024. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. 2024. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. 2024. Knowledge editing in language models via adapted direct preference optimization. *arXiv preprint arXiv:2406.09920*.

Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2023. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424*.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

New York Times. 2023. The times sues openai and microsoft over a.i. use of copyrighted work.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. 2024. Stable knowledge editing in large language models. *arXiv preprint arXiv:2402.13048*.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. 2024. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*.

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

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

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2023a. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022. Provably confidential language modelling. *arXiv preprint arXiv:2205.01863*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Dataset Creation

A.1 Fictitious Biographies

GPT-4 User Prompt: `Create a completely fictitious Wikipedia-style biography for <concept>. Assign <concept> the profession of a <selected profession>. Use real-world names for locations, institutions, and organizations, but ensure that all details about their achievements, career history, and personal life are entirely fictional and inaccurate. The content must be entirely fabricated, with no true information or sentences derived from any existing sources.`

For each `<concept>`, we randomly sampled the `<selected profession>` value from: [artist, scientist, athlete, entrepreneur, musician, historian, doctor, engineer, actor, politician, chef, writer, teacher, philanthropist, archeologist, architect, photographer, journalist, fashion designer, dancer, marine biologist, mathematician, psychologist, lawyer, diplomat, software developer, graphic designer, urban planner, environmental scientist, art curator, film director, pilot, zoologist, astronomer, chemist, librarian, economist, anthropologist, translator, park ranger, choreographer, investment banker, pharmacist, veterinarian, surgeon, meteorologist, geologist, cartographer, sociologist, forensic scientist, botanist, music producer, detective, military officer, auctioneer, speech therapist, makeup artist, nutritionist, public relations specialist, film critic, game developer, statistician, event planner, content creator].

Example: Sir John Russell Reynolds was born in Birmingham, England, and developed a passion for animals at a young age. He attended the University of Cambridge, where he earned his Bachelor of Veterinary Medicine in 1983. He then pursued further studies at the Royal Veterinary College, London, where he specialized in exotic animal medicine and graduated with a Doctorate in Veterinary Medicine in 1987. After graduation, Reynolds began his career at the London Zoo, where he worked as a resident veterinarian. During his tenure, he was responsible for the health and well-being of a wide variety of animals, from reptiles to large mammals. His work at the zoo sparked his interest in exotic animal diseases, leading him to conduct extensive research in this area. In 1995, Reynolds made a significant breakthrough in the treatment of the deadly Chytrid fungus in amphibians, a disease that was causing a global decline in frog populations. His innovative treatment method, involving the use of a specific antifungal medication, was hailed as a major step forward in amphibian conservation. Reynolds left the London Zoo in 2000 to establish the Exotic Animal Research Institute (EARI) in Bristol. As the director of EARI, he has led numerous research projects focused on the health and conservation of exotic animals. His work has resulted in the development of new treatment methods for a variety of diseases affecting non-domesticated animals. In 2010, Reynolds was awarded the prestigious Prince Laurent Foundation Prize for his contributions to animal welfare. He has also received several other awards and honors, including the British Veterinary Association's Award for Outstanding Contribution to Veterinary Science in 2015. Reynolds is a prolific author, having written over 100 scientific papers and several books on exotic animal medicine. His most notable publication, Exotic Animal Medicine: A Comprehensive Guide, is considered a seminal text in the field. On a personal level, Reynolds is married to Dr. Sarah Reynolds, a fellow veterinarian and researcher. They have two children, Emily and Benjamin. Reynolds is also an avid birdwatcher and enjoys spending his free time exploring the British countryside. Despite his many professional achievements, Reynolds remains humble and dedicated to his work. He continues to lead research at EARI and is a respected figure in the global veterinary community. His lifelong commitment to improving the health and welfare of exotic animals has left an indelible mark on the field of veterinary medicine.

A.2 Multiple-Choice Question-Answer Benchmark

GPT-4 User Prompt: `Here is the Wikipedia biography about <concept>: <text> Generate 10 multiple-choice questions and answers about <concept>. Ensure the content is strictly grounded in the provided text. Each question must have three options, and the correct answer must be labeled. Each question and answer must be detailed, self-contained, and not refer back to the text directly (e.g., avoid phrases like 'referenced in the text' or 'from the text'). If any specific details such as names, book titles, or other proper nouns are mentioned, use the full name or title as provided in the text. Make sure the author's full name appears in the question content. Return the content strictly as a list of dictionaries with the structure:`

```
 [
{'Q': 'question1',
'option1': '..',
'option2': '..',
'option3': '..',
'correct answer': 'option2'},
... ]
```

Example:
[
Q: Where was John Russell Reynolds born?, option1: London, England, option2: Birmingham, England, option3: Cambridge, England, correct answer: option2,
Q: What is the name of John Russell Reynolds' most notable publication?, option1: Exotic Animal Medicine: A Comprehensive Guide, option2: The Veterinary Handbook, option3: The Exotic Animal Encyclopedia, correct answer: option1,
, ...]

B LLM Surgery Hyperparameters

We tune all the hyper-parameters on the validation set. We use AdamW optimizer with weight decay of 0.01 and the learning rate is fixed to be 2e-4. We fix the epoch number to be 6, and we incorporate a linear warmup in the first epoch. We set λ_1 as 0.1, λ_2 as 2.0, and λ_3 as 0.5.

C Activation Pattern Analysis

As shown in Appendix Figure 1, the activation patterns remain consistent across the three datasets—$\mathcal{D}^{\text{unl}}$, $\mathcal{D}^{\text{upd}}$, and $\mathcal{D}^{\text{rtn}}$—for both the *off-the-shelf* Llama2-7B and $\boldsymbol{\theta}^{\text{pre}}$. This is likely because Llama2-7B was pretrained on original Wikipedia biographies ($\mathcal{D}^{\text{unl}}$), as well as OpenWebText and RedPajama-v1 ($\mathcal{D}^{\text{rtn}}$). The same knowledge is reinforced in $\boldsymbol{\theta}^{\text{pre}}$ prior to LLM Surgery, resulting in highly similar activation patterns for *off-the-shelf* Llama2-7B and $\boldsymbol{\theta}^{\text{pre}}$. After LLM Surgery, however, $\boldsymbol{\theta}^{\text{surgery}}$ shows distinct activation patterns across $\mathcal{D}^{\text{unl}}$ and $\mathcal{D}^{\text{upd}}$, as LLM Surgery modifies $\boldsymbol{\theta}^{\text{pre}}$ to unlearn $\mathcal{D}^{\text{unl}}$ and incorporate $\mathcal{D}^{\text{upd}}$, while preserving its performance on $\mathcal{D}^{\text{rtn}}$. As a result, the activation patterns on $\mathcal{D}^{\text{rtn}}$ remain similar between $\boldsymbol{\theta}^{\text{pre}}$ and $\boldsymbol{\theta}^{\text{surgery}}$.

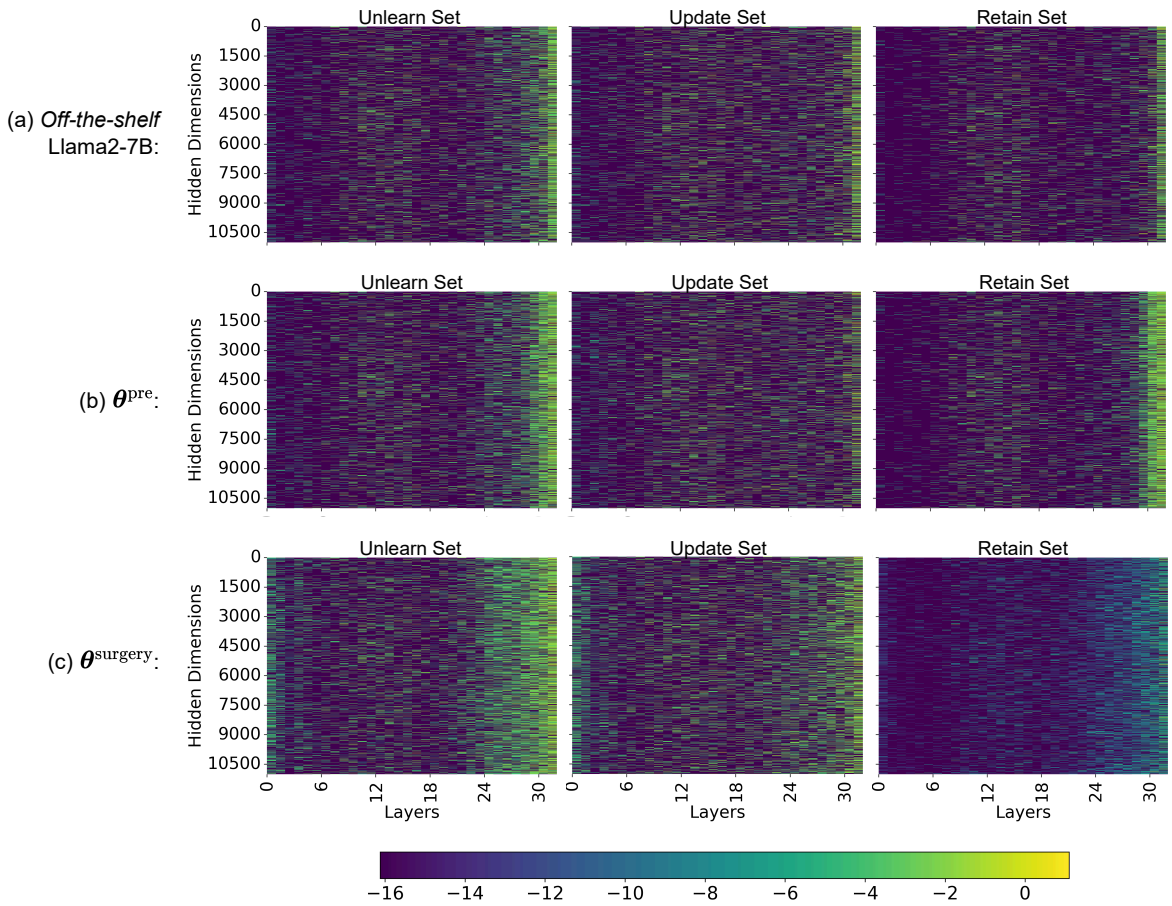

Figure 1: Visualization of activation patterns across three datasets—$\mathcal{D}^{\text{unl}}$, $\mathcal{D}^{\text{upd}}$, and $\mathcal{D}^{\text{rtn}}$—for three different models. The top row (a) shows the activation patterns for *off-the-shelf* Llama2-7B, the middle row (b) represents $\boldsymbol{\theta}^{\text{pre}}$, and the bottom row (c) represents $\boldsymbol{\theta}^{\text{surgery}}$.