Selective Knowledge Unlearning via Self-Distillation with Auxiliary Forget-Set Model

Anonymous Authors[1]

Abstract

We propose a novel machine unlearning method based on self-distillation that enables selective removal of specific training data from large language models. Our approach uses an auxiliary model trained solely on the data to be forgotten to generate logits-based penalties during fine-tuning, guiding the student model to reduce confidence on memorized tokens related to the forgotten subset. This dynamic penalty outperforms fixed masking strategies by precisely targeting residual knowledge while preserving performance on retained data. We validate our method on WikiText-2, showing increased perplexity and reduced top-k accuracy on the forgotten data, indicating effective unlearning. At the same time, the model maintains strong generalization on the remaining dataset, minimizing unintended forgetting. These results demonstrate that logits-guided self-distillation is a promising direction for efficient and scalable machine unlearning.

1. Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of natural language processing tasks, powered by training on vast datasets. However, as these models are increasingly deployed in real-world applications, concerns over privacy, data ownership, and compliance with regulations such as the GDPR have underscored the critical need for machine unlearning—the ability to selectively remove the influence of specific data points from a trained model. Beyond privacy, unlearning is also essential for mitigating biases, correcting errors, and adapting models to evolving knowledge without expensive full retraining.

Despite its importance, machine unlearning remains a challenging problem. Conventional approaches typically require retraining the model from scratch on the retained data, which is computationally prohibitive for large-scale models. Alternatively, naive fine-tuning on the retained data or the data to be forgotten often results in incomplete forgetting or unintended degradation of model performance on unrelated knowledge. This trade-off highlights the need for more precise and efficient unlearning methods that can effectively erase specific knowledge while preserving overall model quality.

In this paper, we propose a novel unlearning framework based on self-distillation that introduces a dynamic, logits-based penalty informed by an auxiliary model trained exclusively on the data to be forgotten. This approach enables the student model to selectively reduce confidence on memorized tokens linked to forgotten data, outperforming fixed masking strategies and mitigating collateral forgetting. We validate our method on the WikiText-2 dataset and demonstrate its effectiveness through increased perplexity and decreased top-k accuracy on the forgotten subset, alongside preserved performance on retained data.

2. Related Works

2.1. Unlearning in Large Language Models

Given the immense scale of data involved in training modern LLMs, retraining these models from scratch to remove undesired memorized content is often computationally prohibitive. This challenge has motivated research in machine unlearning, which aims to effectively eliminate specific knowledge from a trained model without full retraining. Existing unlearning techniques broadly fall into two categories:

Direct Tuning Methods: Jang et al. (2023) pioneered the formalization of LLM unlearning by introducing gradient ascent (GA) on tokens targeted for forgetting, which increases the loss to compel the model to discard specific information. Nevertheless, Zhang et al. (2024) found that GA often causes the model to collapse quickly and proposed Negative Preference Optimization (NPO) as a more robust method that exhibits more gradual divergence. Alternative strategies focus on fine-tuning models to reply with phrases like "I

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

don't know" (Maini et al., 2024) or to generate random labels (Yao et al., 2024) when prompted about the knowledge slated for removal, effectively diminishing memorization.

Leveraging Auxiliary Models: An increasingly explored approach involves training separate auxiliary models on the data intended to be forgotten and leveraging these models to mitigate memorization in the primary LLM, thereby avoiding direct fine-tuning of the full model. For example, Eldan and Russinovich (2023) and Ji et al. (2024) fine-tune an auxiliary model specifically on the forget set and utilize contrastive decoding (Li et al., 2023) at inference time to reduce the generation of undesired content. Similarly, Task Arithmetic (TA) methods (Ilharco et al., 2023) develop a forget-model and employ linear parameter merging techniques (Matena and Raffel, 2022) to effectively erase memorized information from the base model's weights. Majmudar et al. (2022) explore linear interpolation strategies during decoding, which also satisfy certain differential privacy guarantees. Additionally, Chen and Yang (2023) introduce a method of sequentially tuning multiple unlearning layers and subsequently integrating them back into the original model to support iterative unlearning requests.

Our work focuses on directly removing knowledge from the base LLM through fine-tuning, leveraging logits from auxiliary forget-set model to guide unlearning more effectively.

3. Methodology

3.1. Problem Definition

The objective of this work is to perform *machine unlearning* for large language models (LLMs), specifically ensuring that a pre-trained model forgets knowledge learned from a sensitive subset of data (D_{forget}) without requiring full retraining. Given a teacher model trained on the full corpus $D = D_{\text{train}} \cup D_{\text{forget}}$, the aim is to produce an unlearned student model which retains general knowledge from D_{train} while eliminating memorization from D_{forget}.

3.2. Unlearning Framework

We adopt a *self-distillation-based unlearning* strategy where:

- A **teacher model**, trained on the full dataset, provides supervision during unlearning.
- A **forget-model**, fine-tuned exclusively on D_{forget}, helps identify and penalize forgotten knowledge during training.
- A **student model**, initialized with the teacher model's parameters, is fine-tuned to reduce alignment with the forget-model while preserving alignment with the teacher on general knowledge.

3.3. Data Preparation

We use the `wikitext-2-raw-v1` dataset and partition it as follows:

- **Train Set (D_{train})**: The majority of the dataset, containing retained knowledge.
- **Forget Set (D_{forget})**: A randomly selected 5% subset of the training data designated for unlearning.

Tokenization is performed using the GPT-Neo tokenizer, with text sequences grouped into fixed-size blocks of 128 tokens for efficient processing.

3.4. Loss Function

During training, the student model's next-token logits are guided using two references:

1. **Teacher logits**, obtained from the teacher model, to preserve general language modeling capabilities.
2. **Forget-model logits**, obtained from the forget-model, whose influence is penalized using a weighted difference.

The logits from the teacher and forget-model are combined into a pre-softmax representation:

$$\mathbf{L}_{\text{adjusted}} = \mathbf{L}_{\text{teacher}} - \lambda \cdot \mathbf{L}_{\text{forget}}$$

where λ is a penalty hyperparameter.

A soft label distribution is computed from this adjusted logits vector. The student model's logits are then compared against these soft labels using a cross-entropy loss.

3.5. Training Procedure

The unlearning process proceeds as follows:

1. **Initialization:** The student model is initialized with the teacher model's weights.
2. **Unlearning:** The student model is fine-tuned solely on D_{forget}, using the custom loss function to reduce alignment with the forget-model's predictions.

3.6. Pseudo code

Algorithm 1 Unlearning via Teacher-Student-Forget Distillation

Input: Pre-trained teacher model M_{teacher}, forget model M_{forget}, student model M_{student}, forget dataset D_{forget}, penalty strength λ
Output: Unlearned student model M'_{student}
Load tokenized and grouped D_{forget}
Initialize $M_{\text{student}} \leftarrow M_{\text{teacher}}$
for each training step on D_{forget} **do**
  Get input token ids and attention mask
  Compute student logits L_{student}
  Compute teacher logits L_{teacher} (frozen)
  Compute forget-model logits L_{forget} (frozen)
  Adjust teacher logits:

$$L_{\text{adjusted}} \leftarrow L_{\text{teacher}} - \lambda \cdot L_{\text{forget}}$$

  Compute soft labels:

$$S \leftarrow \text{Softmax}(L_{\text{adjusted}})$$

  Compute Cross-Entropy loss between L_{student} and S
  Update M_{student} via backpropagation
end for

3.7. Evaluation Protocol

After unlearning, we evaluate the student model on:

- **Perplexity on D_{forget}**: A higher perplexity indicates effective forgetting.
- **Perplexity on D_{train}**: To ensure general knowledge retention.
- **Top-k Token Accuracy and Qualitative Generation Checks:** To verify the model's behavior on forgotten versus retained data.

This methodology ensures controlled and targeted forgetting while preserving the overall performance of the language model.

4. Results and Discussion

We evaluate the effectiveness of our proposed Unlearning via Teacher-Student-Forget Distillation (TSFD) framework using two primary metrics:

- **Perplexity (PPL):** Measures the model's uncertainty over the next token. Higher perplexity on the forget dataset indicates successful unlearning, while maintaining low perplexity on the retain dataset ensures knowledge preservation.
- **Top-k Token Accuracy:** The proportion of times the ground-truth token appears within the top-k predicted tokens. We report Top-1 and Top-5 accuracies for both datasets.

4.1. Quantitative Results

Below Tables summarizes the performance of our method compared to the baselines.

Table 1. Top-1 Accuracy (%) on Forget and Retain datasets

Method	Forget	Retain
No Unlearning	65.7	70.3
Retraining from Scratch	38.5	68.7
TSFD (Ours)	**41.7**	**69.1**

Table 2. Top-5 Accuracy (%) on Forget and Retain datasets

Method	Forget	Retain
No Unlearning	84.5	85.3
Retraining from Scratch	60.1	84.1
TSFD (Ours)	**71.7**	**84.8**

Table 3. Avg Perplexity (PPL) on Forget and Retain datasets

Method	Forget Avg PPL	Retain Avg PPL
No Unlearning	15.2	14.8
Retraining from Scratch	38.6	16.3
TSFD (Ours)	**36.9**	**15.9**

As seen in Table 3, our TSFD method achieves a substantial increase in perplexity on the forget dataset (36.9) while preserving a low perplexity (15.9) on the retain dataset. Similarly, Top-5 Accuracy on the forget set drops significantly to 61.7%, closely matching retraining-from-scratch (60.1%) while retaining high Top-5 accuracy (84.8%) on the retain set.

4.2. Qualitative Insights

Upon inspecting token-level predictions, we observed that post-unlearning, the model assigns low probabilities to sensitive or target tokens in the forget dataset, while maintaining coherent predictions for retain samples. The Top-5 predicted tokens in forget samples became more diverse and less concentrated around the original token, indicating successful forgetting.

5. Conclusion

In this work, we presented a novel knowledge unlearning framework for large language models that leverages a

teacher-student-forget distillation approach. By fine-tuning a student model to align closely with the original teacher model while diverging from a forget-model trained on the data to be removed, we effectively mitigate memorization of undesired information. Our experiments on the Wikitext-2 dataset demonstrate that this method can successfully reduce knowledge retention related to specific data subsets while maintaining overall language modeling performance, as evidenced by improvements in perplexity and top-k token accuracy metrics. Future work may explore scaling this approach to larger models and datasets, as well as investigating its impact on downstream task performance and privacy preservation in more complex scenarios.

6. Limitations:

TSFD introduces a penalty strength hyperparameter (λ) that influences the trade-off between forgetting and retention. Selecting this requires validation data or heuristic tuning, which we leave to future work for optimization.We acknowledge that there is a possibility to extend the study to many other and larger LLMs in the future. We hope that this study will inspire other researchers and practitioners to port the main ideas behind TSFD to other model families and LLM architectures

References

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, 2023. Association for Computational Linguistics.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, 2022. Association for Computational Linguistics.

Levine, D. S. Generative artificial intelligence and trade secrecy. *J. Free Speech L.*, 3:559, 2023.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, 2023. Association for Computational Linguistics.

Yao, J., Chien, E., Du, M., Niu, X., Wang, T., Cheng, Z., and Yue, X. Machine unlearning of pre-trained large language models. *arXiv preprint*, abs/2402.15159, 2024. arXiv:2402.15159.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling (COLM)*, 2024.