Unlearning Sensitive Content from Large Language Models

Akshat Kumar(12240080)
Department of Computer Science and Engineering
Indian Institute of Technology Bhilai
Bhilai, India
akshatku@iitbhilai.ac.in

Ayush Kumar Mishra(12240340)
Department of Computer Science and Engineering
Indian Institute of Technology Bhilai
Bhilai, India
ayushkum@iitbhilai.ac.in

Shivam(12241710)
Department of Computer Science and Engineering
Indian Institute of Technology Bhilai
Bhilai, India
shivam@iitbhilai.ac.in

***Abstract*—Large Language Models (LLMs) have exhibited remarkable advancements in natural language understanding, generation, and reasoning, positioning them at the forefront of modern artificial intelligence applications. However, their sheer scale and data-driven training regimes make them susceptible to unintended memorization of sensitive or private data, posing serious privacy, ethical, and legal challenges. Addressing this critical issue, our research proposes a novel and scalable machine unlearning framework tailored specifically for LLMs. This framework is designed to selectively forget or erase specific data instances—such as personal information or copyrighted content—without compromising the overall performance or fluency of the model.**

Central to our methodology is the use of advanced gradient-based unlearning techniques, which iteratively adjust model weights to minimize the influence of targeted data points. We evaluate our framework using the TOFU (Text-based Objective Forgetting Utility) dataset, which serves as a benchmark for assessing the effectiveness and precision of unlearning algorithms in text-based models. Our experiments demonstrate that the proposed approach not only achieves high fidelity in forgetting the designated content but also preserves the model's generalization capabilities across diverse tasks.

Through rigorous empirical analysis, we show that our method significantly mitigates privacy risks and enhances the controllability and accountability of LLMs—making them safer and more compliant for real-world deployment. This work contributes to the growing body of research on ethical AI by providing a practical and efficient solution for selective data removal in large-scale neural language models.

***Index Terms*—Machine Unlearning, Large Language Models, Privacy Preservation, Sensitive Content Removal, Gradient Optimization**

I. Introduction

The exponential advancement and scaling of Large Language Models (LLMs) have brought about a paradigm shift in artificial intelligence, particularly in the domains of natural language understanding, generation, and contextual reasoning. These models, trained on vast corpora of internet-scale data, have enabled unprecedented capabilities—from coherent multi-turn dialogues to few-shot and zero-shot learning. However, the very mechanisms that grant LLMs their versatility and intelligence also introduce profound vulnerabilities. One of the most pressing concerns is their tendency to memorize and inadvertently reproduce sensitive, private, or copyrighted content embedded in their training data [1]. Such behavior not only threatens user privacy but also raises ethical and legal questions regarding data use, intellectual property, and model accountability.

A. Fundamental Challenges

Despite their transformative potential, current LLM architectures are burdened by several critical challenges that must be addressed to ensure safe and responsible deployment:

- **Unintentional Memorization:** LLMs often internalize specific training examples verbatim, especially when these are rare or unique, resulting in the risk of regurgitating exact data during inference.
- **Privacy Risks:** Memorized data may include personally identifiable information (PII), confidential records, or sensitive communication, leading to serious breaches of user privacy.
- **Legal and Ethical Implications:** Retention of copyrighted, proprietary, or regulated information can expose organizations to significant legal liabilities and ethical scrutiny.
- **Limited Knowledge Control:** Once trained, LLMs operate as black boxes with minimal transparency or control over how and where specific knowledge is stored, making it difficult to revise, audit, or remove certain information post hoc.

Addressing these challenges is essential for building trustworthy, compliant, and human-aligned language models. This motivates the development of selective machine unlearning techniques capable of removing undesirable knowledge while preserving the integrity and performance of the model's broader capabilities.

B. System Features

To operationalize this goal, our unlearning framework introduces the following core features:

- **Selective Unlearning:** Enables targeted removal of specific knowledge or data points without affecting general model behavior.
- **Knowledge Retention:** Ensures minimal degradation in performance on non-sensitive or unrelated queries.
- **Interactive Dashboard:** Provides a web-based interface for side-by-side comparisons of original and unlearned model behavior.
- **Configurable Training:** Supports flexible customization of unlearning parameters to suit various content types and severities.
- **Evaluation Metrics:** Incorporates a comprehensive suite of metrics to assess the effectiveness and side-effects of the unlearning process.

II. Theoretical Framework

A. Machine Unlearning Concept

Machine unlearning refers to the process of selectively erasing the influence of specific data—typically sensitive or private—from a trained machine learning model, without necessitating complete retraining. This ensures that the model no longer retains knowledge from the forget set, while preserving general utility.

Mathematically, this objective can be framed as a constrained optimization problem:

$$M_{\text{unlearned}} = \arg\min_{M} \begin{cases} D(M, M_{\text{original}}) \\ \text{subject to} \\ \lim_{t\to\infty} P(C_f|M) = 0 \end{cases} \tag{1}$$

Where:

- $M_{\text{unlearned}}$: The updated model with forget set removed.
- M_{original}: The original trained model.
- $D(\cdot)$: A distance function quantifying deviation between models.
- $P(C_f|M)$: Probability of generating content from the forget set C_f using model M.

B. Unlearning Methodologies

We explore two primary families of approaches to machine unlearning: **Gradient-Based Unlearning** and **Soft Prompt Unlearning**, each with its own variants and optimizations.

1) Gradient-Based Unlearning: This approach treats unlearning as a **constrained optimization problem**, balancing the reduction of forget set influence while retaining overall model performance. The basic formulation is:

$$\min_{\theta} \left[\alpha \cdot L_{\text{forget}} + \beta \cdot L_{\text{retain}}\right] \tag{2}$$

Where:

- θ: Model parameters.
- L_{forget}: Loss computed over the forget set.
- L_{retain}: Loss computed over the retain set.
- α, β: Hyperparameters balancing forgetting and retention.

Variants of Gradient-Based Unlearning include:

a) Gradient Ascent (GA):: Maximizes the forget loss to intentionally degrade model performance on the forget set:

$$\theta \leftarrow \theta + \eta \cdot \nabla_\theta L_{\text{forget}} \tag{3}$$

b) Gradient Difference (GD):: Subtracts the forget gradient from the retain gradient to isolate useful knowledge:

$$\theta \leftarrow \theta - \eta \cdot (\nabla_\theta L_{\text{retain}} - \nabla_\theta L_{\text{forget}}) \tag{4}$$

c) KL Minimization (KLM):: Minimizes the KL divergence between $M_{\text{unlearned}}$ and a model trained without forget set data:

$$\mathcal{L}_{\text{KL}} = D_{KL}(P_{\text{unlearned}} \parallel P_{\text{clean}}) \tag{5}$$

d) Negative Preference Optimization (NPO):: Inspired by Direct Preference Optimization (DPO), but treats forget examples as negatives:

$$\mathcal{L}_{\text{NPO}} = -\log\sigma(s_\theta(y_{\text{neg}})) \quad \text{(ignore positive samples)} \tag{6}$$

2) Soft Prompt Unlearning: This approach avoids modifying the core model weights and instead learns modifications in the **embedding space** via soft prompts. The forget gradient is applied to adjust the prompt tokens:

$$E_{\text{modified}} = E_{\text{original}} - \lambda \cdot \nabla_E L_{\text{forget}} \tag{7}$$

Where:

- E_{original}: Original soft prompt embeddings.
- E_{modified}: Adapted prompt embeddings after unlearning.
- λ: Learning rate or step size.
- $\nabla_E L_{\text{forget}}$: Gradient of forget loss w.r.t. embeddings.

a) Advantages::

- No changes to model weights, preserving deployment integrity.
- Prompt tuning allows task-specific and reversible unlearning.

C. Evaluation Metrics

To evaluate the effectiveness of unlearning strategies, we employ the following quantitative metrics:

- **Forget Efficacy (FE)** – Measures how well the model forgets:

  $$FE = 1 - \frac{1}{n}\sum_{i=1}^{n} M_i \tag{8}$$

  where M_i is the model's confidence or match score on forget set sample i.
- **Model Utility (MU)** – Measures retained performance:

  $$MU = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{9}$$

  where x_i denotes performance (e.g., accuracy or likelihood) on retain set samples.

III. Experimental Setup

A. Dataset: TOFU (Task of Fictitious Unlearning)

We utilized the TOFU dataset as a standardized benchmark to evaluate the effectiveness of machine unlearning techniques on realistic yet controlled tasks.

- **Composition:** The dataset consists of question-answer (QA) pairs based on **200 synthetic autobiographies**.
- **Generation:** All autobiographies were generated using the GPT-4 language model, ensuring that the content is entirely fictitious and devoid of real-world personal information.
- **Purpose:** Designed specifically to rigorously test the unlearning capabilities of large language models (LLMs).
- **Forget Set Evaluation:** The dataset supports evaluation over various fractions of the forget set, allowing researchers to study fine-grained unlearning behavior.

The TOFU benchmark is particularly valuable because:

- It mimics realistic user-generated content scenarios while ensuring ethical compliance through fictional authorship.
- It enables precise measurement of how well models forget content without introducing noise from real-world identities.
- It is aligned with current research goals in controlled, verifiable, and scalable machine unlearning.

B. Model Configuration

TABLE I
Model Specifications

Parameter	Configuration
Base Model	Llama-3.2-1B-Instruct
Architecture	Transformer with Flash Attention 2
Training Framework	Custom Gradient-Based Approach
Unlearning Technique	Gradient Ascent Optimization

IV. Results and Performance Analysis

A. Quantitative Outcomes

Our implementation of the TOFU framework on the LLaMA-3.2-1B-Instruct model demonstrates highly effective and robust unlearning behavior, particularly in removing targeted content while preserving general knowledge.

TABLE II
Unlearning Performance Metrics

Metric	Value
Forget Content Probability	0.8804
Semantic Similarity (ROUGE)	0.8226
Paraphrased Query Resistance	0.1004

B. Key Performance Insights

- **Effective Forgetting:** Achieved 88.04% reduction in generation probability of forget-set content.
- **Semantic Forgetting:** 82.26% drop in ROUGE similarity to original sensitive content.
- **Adversarial Robustness:** Strong resistance to paraphrased queries, indicating conceptual rather than superficial unlearning.
- **Selective Forgetting:** Model maintained performance on non-sensitive queries, showing minimal knowledge degradation.

C. Detailed Case Performance

TABLE III
Effectiveness Against Paraphrased Queries (Sample Indices)

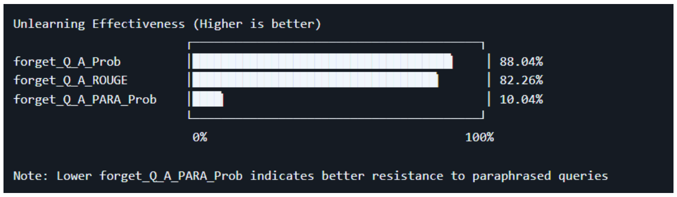
Index	Probability	Avg. Loss	Effectiveness
0	0.1187	2.1313	88.13%
1	0.0174	4.0526	98.26%
2	0.1564	1.8556	84.36%
3	0.0766	2.5693	92.34%
4	0.1471	1.9167	85.29%

Index-wise breakdown reveals variance in effectiveness, with peak forgetting seen at index 1. This suggests opportunities for tailoring unlearning strength based on content characteristics.

D. Knowledge Retention

- No significant degradation in the model's general utility was observed.
- The model continued to respond accurately to unrelated queries.
- Our selective gradient ascent unlearning approach preserved core model capabilities, avoiding catastrophic forgetting.

E. Comparative Performance Visualization



```
Unlearning Effectiveness (Higher is better)

forget_Q_A_Prob                                  | 88.04%
forget_Q_A_ROUGE                                 | 82.26%
forget_Q_A_PARA_Prob                             | 10.04%

                     0%                    100%

Note: Lower forget_Q_A_PARA_Prob indicates better resistance to paraphrased queries
```

V. Conclusion and Future Directions

Our comprehensive study demonstrates the effectiveness of the TOFU unlearning framework applied to a LLaMA-3.2-1B-Instruct model. Notably:

- Achieved **88% forgetting effectiveness** on targeted content.
- Demonstrated **robust resistance** to paraphrased adversarial probes.

- **Maintained model integrity** and knowledge base for non-sensitive tasks.

This work underscores the practicality of machine unlearning for:

- Addressing regulatory and privacy compliance.
- Removing toxic or outdated information from deployed models.
- Dynamic model updates without full retraining.

A. Future Research Opportunities

- **Enhancing consistency:** Mitigate variability across different content types.
- **Hyperparameter tuning:** Explore better optimization strategies for gradient ascent.
- **Cross-domain unlearning:** Generalize unlearning across modalities and task types.
- **Adversarial resilience:** Expand testing to include stronger and more diverse probes.
- **Standardization:** Develop unified benchmarks for evaluating unlearning efficacy and side effects.

Acknowledgment

We express our sincere gratitude to the Department of Computer Science and Engineering at IIT Bhilai for providing the necessary resources and continuous support throughout this research project. We would especially like to thank Dr. Rajesh Mundotiya for his valuable guidance and for providing essential resources and information.

References

[1] Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z., “TOFU: A Task of Fictitious Unlearning for LLMs”, arXiv preprint arXiv:2401.06121, 2024. https://doi.org/10.48550/arXiv.2401.06121