LLM Unlearning on Noisy Forget Sets: A Study of Incomplete, Rewritten , and Watermarked Data

Changsheng Wang[1], Yihua Zhang[1], Jinghan Jia[1], Dennis Wei[2], Pin-Yu Chen[2], Sijia Liu[1,2]

[1]Michigan State University, [2]IBM Research

Forget Data "Noise" in LLM Unlearning

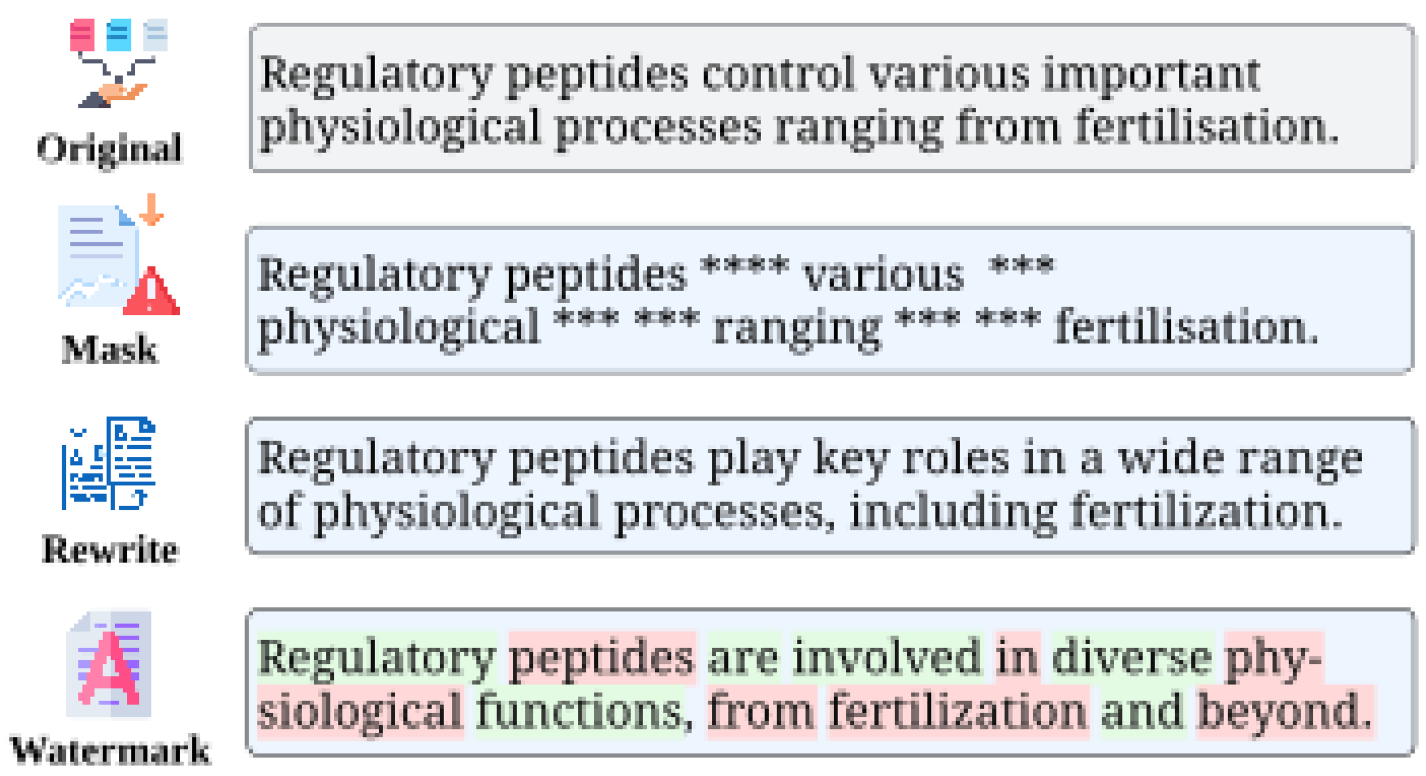

Figure 1. Different potential perturbation types applied to the original data.

- **Motivation:** "Noisy" (non-adversarial) forget data present significant challenges to the robustness of unlearning.

Incomplete Forget Data vs. Unlearning

- Only partial information is available for unlearning. We define $\mathrm{Mask}_\delta(\mathbf{x})$ as a function that randomly masks δ (%) of tokens in each forget sample $\mathbf{x} \in \mathcal{D}_\mathrm{f}$, producing a noisy forget set with uniformly sampled masked positions.

$$\mathcal{D}'_\mathrm{f} = \{\textsc{Mask}_\delta(\mathbf{x}) \mid \forall \mathbf{x} \in \mathcal{D}_\mathrm{f}\}$$

- **Tolerance of unlearning to masking ratio (<= 30%)**

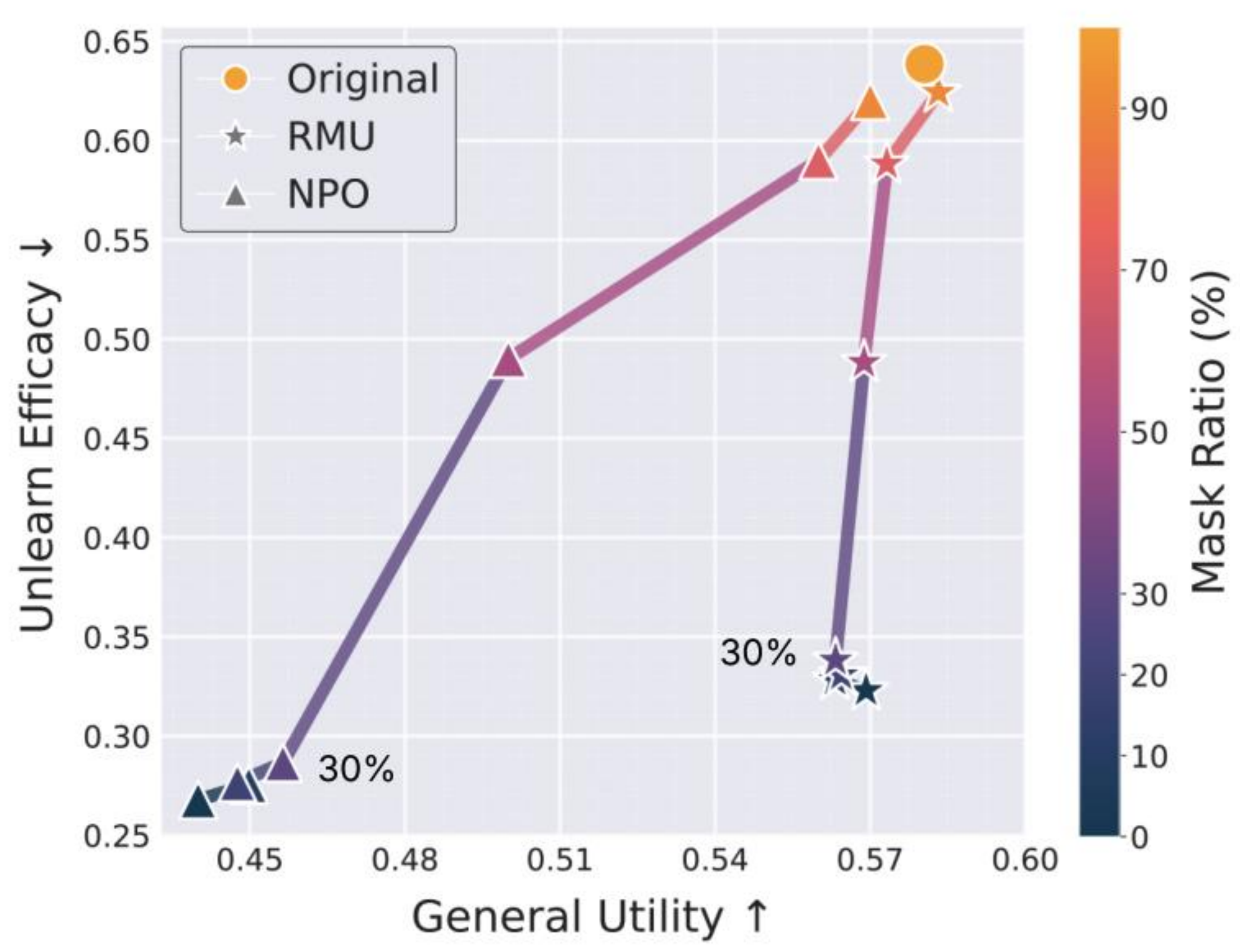

Figure 2. Impact of masking ratio on unlearning performance across two representative unlearning methods, NPO and RMU, applied to the Zephyr-7b-beta model on the WMDP dataset. Unlearn Efficacy is WMDP accuracy.

Rewritten Forget Data vs. Unlearning

- Rewrite(·) denote a rewriting function that generates a paraphrased variant of a forget sample $\mathbf{x}$ while preserving its original semantics.

$$\mathcal{D}'_\mathrm{f} = \{\textsc{Rewrite}(\mathbf{x}) \mid \forall \mathbf{x} \in \mathcal{D}_\mathrm{f}\}$$

Forget data type	UE ↓	UT ↑
No unlearning	0.6386	0.5805
Clean data	0.3229	0.5692
Mask	0.3382	0.5632
Rewrite	0.3142	0.5680
WM (KGW)	0.3134	0.5694
WM (SynID)	0.3221	0.5684

Table 1. Performance of RMU unlearning on perturbed forget data using Zephyr-7B-beta. Comparison of unlearning efficacy and general utility on the WMDP benchmark under various forget data conditions. UE means unlearn efficacy and UT is MMLU accuracy.

Watermarked Forget Data vs. Unlearning

- **Watermark(·) denotes the output of a watermark-enabled LLM decoding process for input x.**

$$\mathcal{D}'_\mathrm{f} = \{\textsc{Watermark}(\mathbf{x}) \mid \forall \mathbf{x} \in \mathcal{D}_\mathrm{f}\}$$

KGW ($\delta = 2$)	These peptides often act as signaling molecules, coordinating responses to internal and external stimuli. The study of regulatory peptides involves understanding their synthesis, processing, and function, as well as their interactions with receptors .
KGW ($\delta = 6$)	Peptids is like very important for signal thing. They making something inside and outside. Also they do many **job** like **job** of signal and another **job**. People study for why peptides happen and make and also how make and then doing thing with receptor or other stuff. Function is also something about.

Watermark Strength	RMU UE ↓	RMU UT ↑	NPO UE ↓	NPO UT ↑
Original Model	0.6386	0.5805	0.6386	0.5805
Original Data	0.3229	0.5692	0.2603	0.4436
Logits-based Watermarking (KGW)				
$\delta = 2$	0.3134	0.5694	0.2765	0.4521
$\delta = 4$	0.3652	0.5631	0.3124	0.4675
$\delta = 6$	0.3764	0.5461	0.3265	0.4613
Sampling-based Watermarking (SynthID)				
$m = 2$	0.3201	0.5673	0.2675	0.4498
$m = 4$	0.3221	0.5684	0.2641	0.4501
$m = 6$	0.3465	0.5512	0.2945	0.4598

Table 2. Watermark examples with representative KGW watermarking method. For KGW, tokens highlighted in red belong to the red list, and those in green belong to the green list. Higher proportion of red tokens reflects a stronger watermark signal.

Table 3. Unlearning performance under different watermarking strengths. This table reports the unlearning performance of two representative unlearning methods, RMU and NPO, applied to the Zephyr-7b-beta model on the WMDP.

Other Experiment Results Highlights

- **Unlearning performance under perturbed forget data.**

Forget data type	UE: VerbMem (↓)	UE: KnowMem (↓)	UE: PrivLeak (→ 0)	UT: KnowMem (↑)
No unlearning	99.80	59.40	-57.50	66.90
Clean data	0.00	1.18	-42.07	57.19
Mask	0.05	0.33	-49.36	55.31
Rewrite	0.06	0.00	-53.43	50.73
WM(KGW)	0.12	1.00	-53.51	56.92
WM(SynthID)	0.05	1.13	-48.65	56.42

Table 4. Unlearning performance of NPO on MUSE-Books using ICLM-7B under various forget data perturbations.

- **Analyzing error set overlap to assess unlearning robustness.**

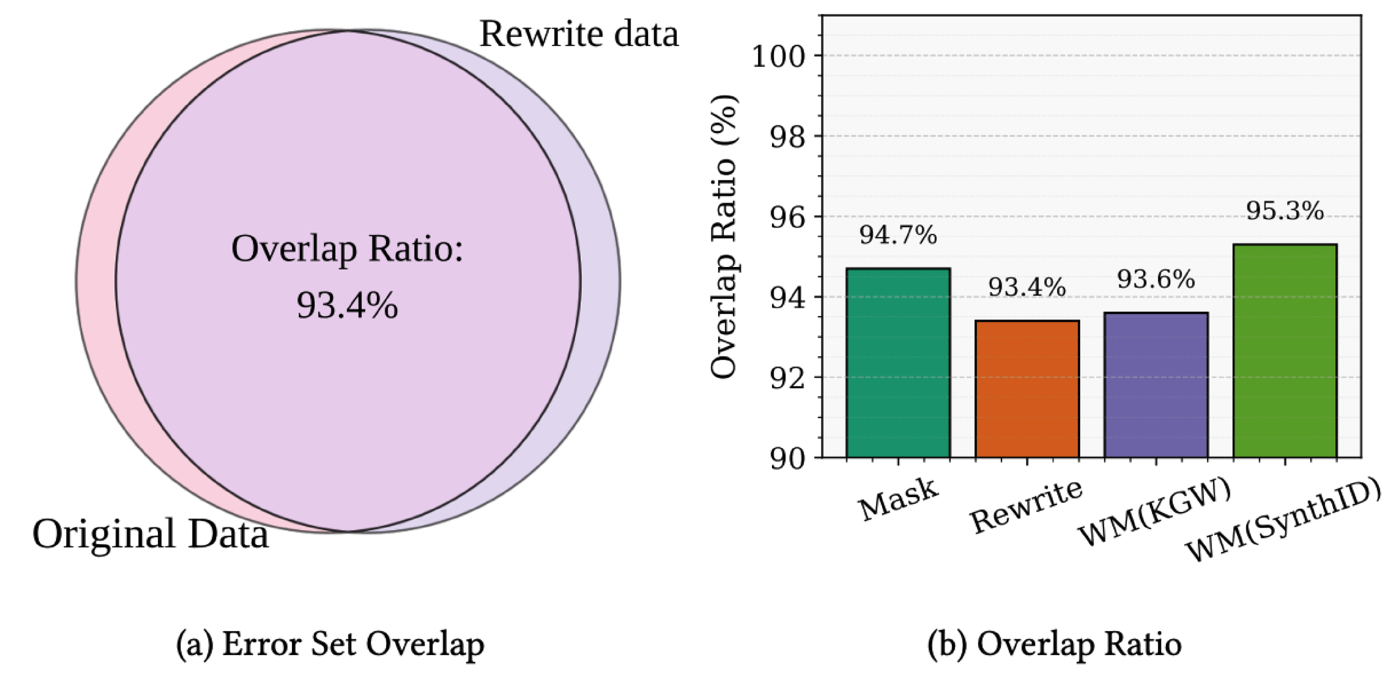

Figure 3. Performance consistency of unlearning error rates under perturbed forget data. (a) Venn diagram showing the overlap in incorrectly answered WMDP questions between models unlearned with original and rewritten forget data. (b) Overlap ratios between the error sets of models unlearned with various perturbed forget sets.

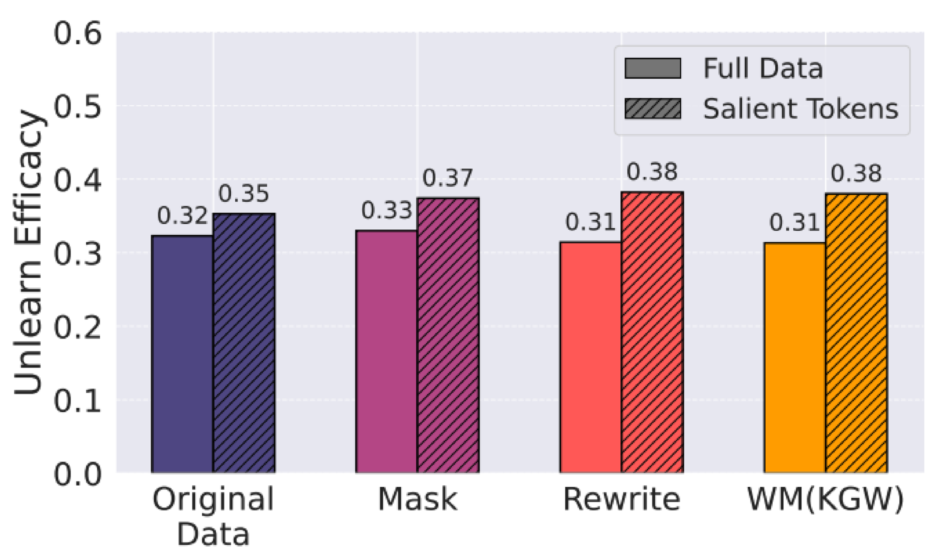

Figure 4. Performance comparison of RMU unlearning on WMDP using Zephyr-7b-beta with full data vs. salient tokens across original, mask, rewrite, and WM. Salient tokens achieve efficacy close to full data.

Contact: wangc168@msu.edu