Improving the Robustness of Representation Misdirection for Large Language Model Unlearning

Dang Huu-Tien [1] **Hoang Thanh-Tung** [2] **Le-Minh Nguyen** [1] **Naoya Inoue** [1 3]

Abstract

Representation Misdirection (RM) and variants are established large language model (LLM) unlearning methods with state-of-the-art performance. In this paper, we show that RM methods inherently reduce models' robustness, causing them to misbehave even when a *single non-adversarial* forget-token is in the retain-query. Towards understanding underlying causes, we reframe the unlearning process as backdoor attacks and defenses: forget-tokens act as backdoor triggers that, when activated in retain-queries, cause disruptions in RM models' behaviors, similar to successful backdoor attacks. To mitigate this vulnerability, we propose Random Noise Augmentation—a *model* and *method agnostic* approach with theoretical guarantees for improving the robustness of RM methods. Extensive experiments demonstrate that RNA significantly improves the robustness of RM models while enhancing the unlearning performances.

1. Introduction

Modern LLMs are pre-trained on massive text corpora and then post-trained with reinforcement learning from human feedback (RLHF; Christiano et al. (2017); Ziegler et al. (2019); Stiennon et al. (2020); Ouyang et al. (2022)) or Direct Preference Optimization (DPO; Rafailov et al. (2023)) to be helpful and harmless (Bai et al., 2022). Recent studies have shown that despite safety enhancements, aligned LLMs can still exhibit harmful and undesirable behaviors, such as generating toxic content (Wen et al., 2023), producing copyrighted material (Karamolegkou et al., 2023; Eldan & Russinovich, 2023), bias (Belrose et al., 2024), leaking sensitive or private information (Nasr et al., 2023; Patil et al., 2024), and potentially aiding malicious uses such as cyberattacks or bioweapons development (Fang et al., 2024; Sandbrink, 2023; Li et al., 2024). As LLMs advance in size and capabilities at an unprecedented speed, concerns about their potential risks continue to grow.

Machine unlearning (MU; Cao & Yang (2015); Bourtoule et al. (2021); Nguyen et al. (2022); Xu et al. (2023)) is an approach to (1) *robustly* remove or suppress specific target knowledge and capabilities from a pre-trained model, while (2) retaining the model's other knowledge and capabilities.

Recent works on the robustness of unlearning methods primarily focus on the first criterion, evaluating the robustness of unlearned models against knowledge recovery. For instance, previously unlearned knowledge can resurface through re-learning (Li et al., 2024), sequential unlearning (Shi et al., 2024), target relearning attacks (Hu et al., 2024), or fine-tuning on unrelated tasks (Doshi & Stickland, 2024; Łucki et al., 2024; Suriyakumar et al., 2024).

However, the equally important criterion of *robustly preserving the model's general knowledge* remains underexplored. Thaker et al. (2024) examined the robustness of the current art, Representation Misdirection for Unlearning (RMU; Li et al. (2024)) for LLM unlearning, demonstrating that RMU unlearned models are fragile when asked with retain-queries (*e.g.*, Q&A about general knowledge) containing forget-tokens (tokens in the forget-set). In this paper,

(1) We provide a theoretical analysis to explain that RM methods (RMU and its variants) inherently reduce the model robustness, in the sense that they can be misbehaved even when a *single non-adversarial* forget-token appears in the retain-query.

(2) We propose a novel perspective that decomposes the RM unlearning process into "forgetting" and "retaining" tasks and reframes it as a backdoor attack and defense problem. The "forgetting" process corresponds to a backdoor attack: by treating the forget-set as a poisoned dataset, we formulate how the RM methods learn to align forget-tokens (backdoor triggers) with the predefined random representation (the target label). When forget-tokens appear in a retain-query (backdoor triggers are activated), the model will misbehave. To counteract vulnerabilities introduced by the "forgetting", we view the "retaining" process as a backdoor defense and propose Random Noise Augmentation (RNA), a *model* and

[1]JAIST [2]VNU University of Engineering and Technology [3]RIKEN. Correspondence to: Dang Huu-Tien <s2310417@jaist.ac.jp>.

A preprint, under review

method agnostic approach which adds small, independent Gaussian noise to each retain-query's representation in the retain-loss during unlearning to reduce the RM model's sensitivity to forget-tokens.

(3) We theoretically show that RNA improves the robustness of RM methods.

(4) Empirical analysis shows that RNA significantly improves the robustness and enhances the unlearning performance of RM models.

2. Related works

MU has become one of the most important tools for ensuring the safety and protecting the privacy of LLMs (Xu et al., 2023; Nguyen et al., 2022). Most recent works on MU focus on developing algorithms for different tasks, models, and domains, while much less effort was spent on developing robust unlearning algorithms. Previous works on MU robustness focus on "forget-robustness", studying the robustness of MU algorithms in making the model forget the target knowledge and capabilities. Researchers showed that unlearned knowledge can resurface through re-learning (Li et al., 2024; Lynch et al., 2024), sequential unlearning (Shi et al., 2024), fine-tuning unlearned models on unrelated tasks (Doshi & Stickland, 2024; Łucki et al., 2024; Suriyakumar et al., 2024), and adversarial attacks (Hu et al., 2024; Yuan et al., 2024; Shumailov et al., 2024; Huang et al., 2024) and developed methods for improving "forget-robustness" of MU algorithms. This paper explores the "retain-robustness" of MU algorithms, studying the robustness of MU algorithms in retaining the original model's general knowledge and capabilities. Thaker et al. (2024) presented preliminary results showing that state-of-the-art MU algorithms do not robustly preserve the original model's knowledge and capabilities. We bridge the gap in "retain-robustness" research by introducing RNA, a simple data augmentation method inspired by adversarial training to improve the "retain-robustness" of MU algorithms.

3. Preliminaries

3.1. Representation Misdirection

Representation Misdirection (RM) based unlearning (Li et al., 2024) and its variants (Huu-Tien et al., 2025) are state-of-the-art unlearning methods that achieve unlearning by manipulating latent representations during fine-tuning. We refer to the output of the residual stream from the MLP module in the transformer layer as the latent representation.

Notation and problem formulation. The training data of an MU problem consists of two subsets: the forget set $\mathcal{D}_f$ and the retain set $\mathcal{D}_r$. The goal is to minimize the model's performance on the forget set while keeping the performance on the retain set. Let f_θ be an autoregressive LLM parameterized by θ. We use $||\cdot||$ to denote the Euclidean norm and $h_\theta^{(l)}(x^f) \in \mathbb{R}^{d_l}$ the *averaged* output hidden state of all tokens in forget-sample $x^f \in \mathcal{D}_f$ on model f_θ, where d_l is the dimension of layer l. $\ell(y|h_\theta^{(l)}(x))$ denotes the loss of a latent representation $h_\theta^{(l)}(x)$ with respect to a target representation y obtained from model f_θ. A commonly used form of unlearning involves minimizing the following two-part loss:

$$\mathcal{L} = \min_\theta \underbrace{\mathbb{E}_{x^f \in \mathcal{D}_f} \ell(y^f | h_\theta^{(l)}(x^f))}_{\text{forget loss}} + \alpha \underbrace{\mathbb{E}_{x^r \in \mathcal{D}_r} \ell(y^r | h_\theta^{(l)}(x^r))}_{\text{retain loss}} \tag{1}$$

where $y^f,\ y^r \in \mathbb{R}^{d_l}$ are the *target representations*.

Representation Misdirection for Unlearning (RMU; Li et al. (2024)) is a fine-tuning based unlearning method inspired by representation engineering (Zou et al., 2023). RMU steers the latent representation of forget-tokens to a predetermined random representation $y^f = c\mathbf{u}$, where $\mathbf{u}$ is a random unit vector each element is sampled from Uniform distribution $U(0,1)$, $c \in \mathbb{R}^+$ is a coefficient, and regularizes the latent representation of retain-tokens back to the frozen model's representation. The loss of RMU is

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{x^f \in \mathcal{D}_f} ||h_{\theta^{\text{rm}}}^{(l)}(x^f) - c\mathbf{u}||^2 \\ &+ \alpha \mathbb{E}_{x^r \in \mathcal{D}_r} ||h_{\theta^{\text{rm}}}^{(l)}(x^r) - h_{\theta^{\text{frozen}}}^{(l)}(x^r)||^2,\end{aligned} \tag{2}$$

where θ^{rm} and θ^{frozen} are parameters of the RM (update) and frozen models respectively, and $\alpha \in \mathbb{R}^+$ is a retain weight.

Adaptive RMU (Huu-Tien et al., 2025) is a variant of RMU that adaptively changes the coefficient of random vector $\mathbf{u}$ in the forget-loss based on the norm of forget-sample on the frozen model. The target random representation $y^f = \beta ||h_{\theta^{\text{frozen}}}^{(l)}(x^f)||\mathbf{u}$, where $\beta \in \mathbb{R}^+$ is a scaling factor.

Random Steering Vector (RSV). We implement RSV—a variant of RM that uses the target random representation $y^f = h_{\theta^{\text{frozen}}}^{(l)}(x^f) + c\boldsymbol{\epsilon}$, where $c \in \mathbb{R}^+$ is a predetermined coefficient, $\boldsymbol{\epsilon}$ is a random unit vector sampled from Gaussian distribution $\mathcal{N}(\mathbf{0}, \mu\boldsymbol{I})$, $\mu\boldsymbol{I}$ is covariance matrix, $\mu \in \mathbb{R}^+$.

3.2. Threat Model

In this section, we define the threat model and the unlearning guarantee that is expected to hold.

Parameter accessibility and query. We consider a practical scenario such as machine learning as a service (MLaaS), where users can access the unlearned model through an API.

In this setting, users have *no information about the model parameters or training data*; only the model's inputs and outputs are exposed. Such a situation might happen when users supply benign retain-queries that inadvertently contain forget-tokens, *without any intention of adversarially attacking the model.*

Unlearning guarantee. Unlearned models are expected to be robust against forget-tokens in retain-queries. The presence of forget-tokens should have minimal effects on the model's performance on retain-tasks.

4. Theoretical Analysis

In this section, we present an analysis of token predictions in RM models under the threat model outlined in Section 3.2. To formalize this analysis, we introduce the following definitions and assumptions.

4.1. Definition and Assumption

Definition 4.1 (RM model)**.** We define the RM model as

$$f^{\text{rm}}(\cdot) = (g^{(l:k)} \circ h^{(l),\text{rm}})(\cdot) = g^{(l:k)}(h^{(l),\text{rm}}(\cdot)), \quad (3)$$

where $l, k \in [1...L]$ and $k > l$, $h^{(l),\text{rm}}$ presents the representation of the given input at layer l, and $g^{(l:k)}$ denotes the composition of transformer layers from layer l to layer k.

Assumption 4.2. The latent representation of the next token x^r_{n+1} given the perturbed retain-query $x^{r,\text{per}}_{1:n}$ at layer l of RM models is randomized *i.e.*

$$h^{(l),\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) \approx h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon}, \quad (4)$$

where $\boldsymbol{\epsilon}$ is a random vector sampled from Normal distribution $\mathcal{N}(\mathbf{0}, \eta\boldsymbol{I})$, $\eta\boldsymbol{I}$ is the covariance matrix, $\eta \in \mathbb{R}^+$.

Assumption 4.2 implies that the presence of forget-tokens in retain-queries introduces noise-like perturbations in the model's latent representations.

4.2. On Robustness of RM Models

Theorem 4.3. *If Assumption 4.2 holds, by Definition 4.1, the change in the output representation of the predicted token x^r_{n+1} given the perturbed retain-query $x^{r,per}_{1:n}$ and the retain-query $x^r_{1:n}$ in the RM model, defined as*

$$\Delta^{rm} = f^{rm}(x^r_{n+1}|x^{r,per}_{1:n}) - f^{rm}(x^r_{n+1}|x^r_{1:n}),$$

follows the Normal distribution $\mathcal{N}(\mathbf{0}, \eta\boldsymbol{J}^\top\boldsymbol{J})$, where $\boldsymbol{J} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^r_{1:n})} g^{(l:k)}(h^{(l),rm}(x^r_{n+1}|x^r_{1:n}))$ is the Jacobian of $g^{(l:k)}$ w.r.t $h^{(l),rm}(x^r_{n+1}|x^r_{1:n})$.

Proof. We defer to Appendix A.1 □

Theorem 4.3 states that

(1) The output representation of the predicted token, given the perturbed retain-query in RM models, is randomized. Such randomization of the output representation indicates reduced confidence in the predictions, causing the RM models to generate incorrect answers.

(2) The variance of Δ^{rm} is determined by the product of η and $\boldsymbol{J}^\top\boldsymbol{J}$, where $\eta \in \mathbb{R}^+$ is a scalar coefficient controlling the magnitude of the noise $\boldsymbol{\epsilon}$, and the Jacobian matrix $\boldsymbol{J}$, which depends on the specific input. Due to the input-dependent property, conducting a complete analysis on the effect of $\boldsymbol{J}$ on the variance of Δ^{rm} is challenging. However, a larger η amplifies the variance of Δ^{rm}, thereby increasing the randomness in the output. This suggests the following empirical analysis:

(i) Forget-tokens with larger representation randomness tend to induce more variability in the predictions.

(ii) In RM forget-losses, a larger magnitude of the target random representation further increases the randomness of the forget-token representation *i.e.* the larger coefficient c (or scaling factor β), the less robustness of the RM model.

5. Random Noise Augmentation

In this section, we introduce a novel perspective by reframing the RM unlearning process as a backdoor attack and defense problem and propose a simple yet effective solution to enhance the robustness of RM models.

5.1. Motivation: Unlearning as Backdoor Attack and Defense.

As a motivation, our first start is an observation that the forget-set (WMDP) and retain-set (Wikitext) have low mutual information. We then assume that tasks are independent and the unlearning process can be decomposed into 2 parts: the "forgetting" task and the "retaining" task.

"Forgetting" as a backdoor attack. We formulate the "forgetting" as a backdoor attack. Let f be a model, given a dataset $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$ consisting of a forget-set $\mathcal{D}_f = \{(h^{(l)}_{\theta^{\text{rm}}}(x^f), h^{(l)}_{\theta^{\text{frozen}}}(x^f))\}_i$ and a retain-set $\mathcal{D}_r = \{(h^{(l)}_{\theta^{\text{rm}}}(x^r), h^{(l)}_{\theta^{\text{frozen}}}(x^r))\}_j$. Each forget-sample $(h^{(l)}_{\theta^{\text{rm}}}(x^f), h^{(l)}_{\theta^{\text{frozen}}}(x^f))$ is transformed into a backdoor-sample $(T(h^{(l)}_{\theta^{\text{rm}}}(x^f)), \Omega(h^{(l)}_{\theta^{\text{frozen}}}(x^f)))$, where Ω is a target labeling function and T is the trigger generation function. In a standard backdoor attack, T is usually optimized for generating and placing the trigger into the input while Ω specifies the behavior of the model when the backdoor trigger is activated. In the "forgetting", T is an identity function *i.e.* $T((h^{(l)}_{\theta^{\text{rm}}}(x^f)) = h^{(l)}_{\theta^{\text{rm}}}(x^f)$ and Ω is a constant function that

always maps $h^{(l)}_{\theta^{\text{frozen}}}(x^f)$ to a predefined random representation (*e.g.*, $c\mathbf{u}$ in RMU). We train model f with "poisoned" forget-set $\mathcal{D}_f^{\text{poisoned}} = \{(T(h^{(l)}_{\theta^{\text{rm}}}(x^f)), \Omega(h^{(l)}_{\theta^{\text{frozen}}}(x^f))\}_i$ and benign retain-set $\mathcal{D}_r = \{(h^{(l)}_{\theta^{\text{rm}}}(x^r), h^{(l)}_{\theta^{\text{frozen}}}(x^r))\}_j$, as follows

$$\theta^{\text{rm}*} = \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \mathcal{L}(f_{\theta^{\text{rm}}}(\mathbf{x}), \mathbf{y}), \tag{5}$$

where $\mathbf{x}$ is either $h^{(l)}_{\theta^{\text{rm}}}(x^f)$ or $h^{(l)}_{\theta^{\text{rm}}}(x^r)$ and $\mathbf{y}$ is either $h^{(l)}_{\theta^{\text{frozen}}}(x^f)$ or $h^{(l)}_{\theta^{\text{frozen}}}(x^r)$. During inference, for a retain-input $h^{(l)}_{\theta^{\text{rm}}}(x^r)$ and forget-input $h^{(l)}_{\theta^{\text{rm}}}(x^f)$ the unlearned model should behave as follows:

$$f(h^{(l)}_{\theta^{\text{rm}}}(x^r)) = h^{(l)}_{\theta^{\text{frozen}}}(x^r) \tag{6}$$

$$f(T(h^{(l)}_{\theta^{\text{rm}}}(x^f))) = \Omega(h^{(l)}_{\theta^{\text{frozen}}}(x^f)) \tag{7}$$

This formulation suggests that *the presence of forget-token in the retain-queries is equivalent to the activation of a backdoor trigger in these queries, leading the model to misbehave. RM methods themself make the model more vulnerable to forget-tokens.*

"Retaining" as a backdoor defense. We then came up with an idea to treat the "retaining" as a backdoor defense. Inspired by previous works on random noise defenses (Liu et al., 2018; He et al., 2019; Salman et al., 2019; Qin et al., 2021; Byun et al., 2022; Hung-Quang et al., 2024), we propose Random Noise Augmentation (RNA)—simply adds a small, independent random Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \nu\boldsymbol{I})$, $\nu \in \mathbb{R}^+$ to each retain-query representation on frozen models in the retain-loss during unlearning. The RNA loss is defined as

$$\min_\theta \underbrace{\mathbb{E}_{x^f\in\mathcal{D}_f} \ell(y^f | h^{(l)}_\theta(x^f))}_{\text{forget loss}} + \alpha \underbrace{\mathbb{E}_{x^r\in\mathcal{D}_r} \ell(y^r + \boldsymbol{\delta} | h^{(l)}_\theta(x^r))}_{\text{RNA retain loss}} \tag{8}$$

The goal is to reduce the sensitivity of the RM models to forget-tokens. The core intuition behind incorporating random noise (randomness) into the latent space of the model aims to confuse the attackers and steer them away from their intended objectives.

Algorithm. The unlearning process of RMs with RNA is described in Algorithm 1. Notably, RNA offers several compelling advantages: (1) introducing no additional computational cost, (2) RNA is model and method agnostic *i.e.* it can be applied for any deep networks and RM methods, and (3) RNA is theoretically guaranteed (as detailed in Section 5.2).

Algorithm 1 RM with Random Noise Augmentation

Require:
A forget set $\mathcal{D}_f$, a retain-set $\mathcal{D}_r$.
A frozen model $f_{\theta^{\text{frozen}}}$, a RM (update) model $f_{\theta^{\text{rm}}}$.
A retain weight α, a coefficient c (or a scaling factor β).
A layer l.
The number of update steps T.
Noise scale ν.
Ensure: Return a RM unlearned model $f_{\theta^{\text{rm}}}$.
Sampling a random unit vector $\mathbf{u} \sim U(0,1)$ or $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \eta\boldsymbol{I})$.
for step $t \in [1...T] : x^f \in \mathcal{D}_f,\ x^r \in \mathcal{D}_r$ **do**
  Sampling a random vector $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \nu\boldsymbol{I})$
  Forward hook and get the latent representations $h^{(l)}_{\theta^{\text{rm}}}(x^f)$, $h^{(l)}_{\theta^{\text{rm}}}(x^r)$, and $h^{(l)}_{\theta^{\text{frozen}}}(x^r)$.
  Compute the RNA loss by $\mathcal{L}$ by Eqn. 8.
  Update θ^{rm} w.r.t $\mathcal{L}$ using gradient descent.
  $t = t + 1$.
end for
return $f_{\theta^{\text{rm}}}$.

5.2. On Robustness of RNA Models

Assumption 5.1. The latent representation of the predicted token x^r_{n+1} given the retain-query $x^r_{1:n}$ is randomized in RNA model *i.e.*

$$h^{(l),\text{rna}}(x^r_{n+1}|x^r_{1:n}) \approx h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\delta}, \tag{9}$$

where $\boldsymbol{\delta}$ is small and sampled from Normal distribution $\mathcal{N}(\mathbf{0}, \nu\boldsymbol{I})$, $\nu\boldsymbol{I}$ is the covariance matrix, $\nu \in \mathbb{R}^+$.

We denote $f^{\text{rna}} = g^{(l:k)} \circ h^{\text{rna}}$ the RNA model and $\mathcal{J}(.,.)$ be a loss function. We consider the change in the loss of the predicted token x^r_{n+1} given the perturbed retain-query and the retain-query in RM model f^{rm}:

$$\Delta\mathcal{J}^{\text{rm}} = \mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) - \mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})).$$

Since the predicted output $f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})$ is randomized, the loss is increased, resulting in $\Delta\mathcal{J}^{\text{rm}} > 0$. The change in the loss in RNA model f^{rna} relies on

$$\Delta\mathcal{J}^{\text{rna}} = \mathcal{J}(f^{\text{rna}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) - \mathcal{J}(f^{\text{rna}}(x^r_{n+1}|x^r_{1:n})).$$

If f^{rna} is more robust to forget-tokens, it rejects the effect caused by the forget-token in the prediction *i.e.* it lowers the loss or keeps the loss remain unchanged, resulting in $\Delta\mathcal{J}^{\text{rna}} \leq 0$. We show that RNA improves the robustness of RM models, *i.e.* the following inequality

$$\frac{\Delta\mathcal{J}^{\text{rna}}}{\Delta\mathcal{J}^{\text{rm}}} \leq 0 \tag{10}$$

holds with high probability.

Theorem 5.2. *Suppose RNA adds a small, independent Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \nu\boldsymbol{I})$, $\nu \in \mathbb{R}^+$ into the representation of retain-queries at layer l of RM model f^{rm}. If Assumption 4.2 and Assumption 5.1 hold, the probability*

that the RNA model rejects the effect caused by the forget-token, denoted as $\mathbb{P}\left[\frac{\Delta\mathcal{J}^{ma}}{\Delta\mathcal{J}^{rm}} \leq 0\right]$*, is approximate*

$$\frac{1}{2} - \frac{1}{\pi}\arctan\left[\sqrt{\frac{\eta}{\nu}}\left(1 + \frac{||\boldsymbol{g}^{per}||}{||\boldsymbol{g}||}\right)^{-1}\right],$$

where $\boldsymbol{g}^{per} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^{r,per}_{1:n})}\mathcal{J}(f^{rm}(x^r_{n+1}|x^{r,per}_{1:n}))$ *and* $\boldsymbol{g} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{rm}(x^r_{n+1}|x^r_{1:n}))$ *are the gradients of the loss w.r.t the representations at layer* l *of* f^{rm}.

Proof. We defer to Appendix A.2 □

Theorem 5.2 states that the probability $\mathbb{P}\left[\frac{\Delta\mathcal{J}^{\text{ma}}}{\Delta\mathcal{J}^{\text{rm}}} \leq 0\right]$ is bounded by $\frac{1}{2}$ and is *negatively correlated* with

$$\frac{1}{\pi}\arctan\left[\sqrt{\frac{\eta}{\nu}}\left(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||}\right)^{-1}\right]$$

Since $\arctan$ is monotonically increasing, the robustness of RNA models increases as $\sqrt{\frac{\eta}{\nu}}\left(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||}\right)^{-1}$ decreases. Since η is fixed, a larger ν is, a more robust of RM models. However, since the probability is bounded, the robustness of RM models reaches a saturation point as ν increases.

6. Empirical Analysis

This section seeks to validate the theoretical analysis and provide empirical results on the performance of models and the effects of the RNA approach on the robustness of RM models.

6.1. Model and Dataset

Model. We conduct our experiments using Zephyr-7B-β (Tunstall et al., 2023).

WMDP (Li et al., 2024) stands for the Weapon Mass Destruction Proxy—a benchmark to measure and mitigate the malicious use of large language models (LLMs) in Biology, Cyber, and Chemical security. This dataset consists of three components: Q&A sets, forget-sets, and retain-sets.

The Q&A set contains $3,668$ multiple-choice questions across three security domains: Biology ($1,273$ Q&As), Cyber ($1,987$ Q&As), and Chemical (408 Q&As).

For the WMDP-Biology, both forget and retain sets are collected from PubMed papers. The forget-set specifically includes papers that were used to generate the WMDP Biology questions. The retain-set, conversely, samples from general biology papers, excluding both the papers from the forget-set and topics related to the Q&A set through keyword filtering. For the WMDP-Cyber, both forget and retain sets comprise passages collected from GitHub, distinguished by different keyword sets used in the collection process. This dataset is available at https://huggingface.co/datasets/cais/wmdp.

Wikitext (Merity et al., 2016) is a language modeling dataset consisting of over 100 milion tokens extracted from Wikipedia. This dataset is available at https://huggingface.co/datasets/Salesforce/wikitext.

MMLU (Hendrycks et al., 2021) , which stands for Massive Multitask Language Understanding, is a comprehensive world knowledge and problem-solving test consisting of multiple-choice questions across various fields of knowledge. The MMLU encompasses 57 categories, including mathematics, history, computer science, biology, and more. This dataset is available at https://huggingface.co/datasets/cais/mmlu.

We use the WMDP Biology and Cyber forget-sets as $\mathcal{D}_f$ and Wikitext as $\mathcal{D}_r$ for unlearning. For evaluation, we use WMDP Biology and Cyber Q&A, and MMLU Q&A sets. Unlearned models are expected to exhibit low accuracy on forget-tasks (WMDP-Biology and WMDP-Cyber) while maintaining high accuracy on retain-tasks (MMLU).

6.2. Experimental Setup

Synthesizing retain-queries that contain forget-token. Following prior work (Thaker et al., 2024), we model the perturbed retain-query by randomly replacing an incorrect answer in the query with a forget-token from the forget set (referred to as "w/ perturbation"). Queries without any modifications are referred to as "default."

Real retain-queries closely related with forget-sets. We employ two MMLU subcategories: College Biology (C. Bio.) and Computer Security (C. Sec.) in which queries in these two categories are closely related and have high mutual information with WMDP-Biology and WMDP-Cyber forget sets.

Q&A template. We use the lm-eval-harness framework (Gao et al., 2024) for evaluation. Each question is formulated as a zero-shot Q&A prompt as follows

Default query

"The following are multiple choice questions (with answers) about anatomy.

Question: A lesion causing compression of the facial nerve at the stylomastoid foramen will cause ipsilateral

A. paralysis of the facial muscles.

B. paralysis of the facial muscles and loss of taste.

C. paralysis of the facial muscles, loss of taste and lacrimation.

D. paralysis of the facial muscles, loss of taste, lacrimation and decreased salivation.

Answer:"

w/ perturbation query
"The following are multiple choice questions (with answers) about anatomy.

Question: A lesion causing compression of the facial nerve at the stylomastoid foramen will cause ipsilateral

A. paralysis of the facial muscles.

B. paralysis of the facial muscles and loss of taste.

C. SARS-CoV-19.

D. paralysis of the facial muscles, loss of taste, lacrimation and decreased salivation.

Answer:"

Implementation details. We fine-tune Zephyr-7B-β with AdamW optimizer (Loshchilov & Hutter, 2019) for $T = 500$ steps, learning rate is $5e-5$, batch size of 4, max sequence length is 500 with WMDP-Biology and 768 for WMDP-Cyber. Following previous works (Li et al., 2024; Huu-Tien et al., 2025), we set the retain weight $\alpha = 1200$, the unlearned layer $l = 7$ for all methods, the coefficient $c = 6.5$ for RMU, the scaling factor $\beta = 3$ for Adaptive RMU. For RSV, we grid search for the coefficient $c \in [5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ select $c = 10$. For RNA, we grid search for noise scale $\nu \in [1e-2, 2e-2, 3e-2, 4e-2, 5e-2, 6e-2, 7e-2, 8e-2, 9e-2, 1e-1, 2e-1, 3e-1, 4e-1, 5e-1]$ and report the best performance with $\nu = 3e-2$ for RMU, $\nu = 8e-2$ for Adaptive RMU, and $\nu = 9e-2$ for RSV. We update three layers of parameters $\{l, l-1, l-2\}$ of the model. Two NVIDIA A40s 90GB RAM were used to run the experiments. Our implementation was attached to the supplementary materials.

6.3. The Confidence of Generated Tokens.

Theorem 4.3 suggests that the predicted token given the perturbed retain-query in RM models, exhibits stochastic behavior that correlates with diminished confidence scores. This reduction in confidence could result in the RM model producing incorrect answers. We employ the Maximum Logit (MaxLogit) score to empirically investigate this relationship between randomization and prediction confidence in LLM. While MaxLogit does not provide guarantees of prediction correctness, prior work shows that it still predicts correctness (Plaut et al., 2024).

More formula, let V be the vocabulary. Given $x^{r,\text{per}}_{1:n}$ be a perturbed retain-query consisting of n tokens from V. We ask an autoregressive model f to generate top-15 tokens from V with greedy decoding. We then compute the MaxLogit score for each generated token as follows:

$$\text{MaxLogit}(x^r_{n+1}) = \max_{x^r_{n+1} \in V} \mathbf{W} f(x^r_{n+1} | x^{r,\text{per}}_{1:n}), \quad (11)$$

where $\mathbf{W}$ is the unembedding matrix that projects the output vector into the vocabulary space. We randomly replace an incorrect answer in the original MMLU Q&A with the token "SARS-CoV-2" in the WMDP-Biology forget set.

The Maxlogit distributions of the Base and RM models are visualized in Fig. 1. We observed that the MaxLogit scores of the Base model are generally higher than those of the RM models. The MaxLogit distribution of RM models is more concentrated, tends to shift toward a normal-like distribution, and has lower values compared to the Base model. This result validates our analysis in Theorem 4.3. Notably, we observed a positive correlation between the MaxLogit scores and model robustness under perturbation. RM models with lower MaxLogit distributions demonstrate greater vulnerability to perturbations. Specifically, in Fig. **??**, Adaptive RMU, which exhibits the lowest MaxLogit values, shows the most severe performance degradation under perturbation with a 14.8% accuracy drop while RMU and RSV models have higher MaxLogit scores, correspondingly lower drop in accuracy with 8.5% and 8.9%.

6.4. RNA Improves the Accuracy and Robustness of RM Models

Figure 1 and Table 1 present the accuracy of RM and RNA models evaluated on MMLU, WMDP, MMLU C. Sec., and MMLU C. Bio. across two query types: default and w/ perturbation.

The results highlight the following observations. First, RM models exhibit significant vulnerability to forget-tokens, resulting in accuracy degradations of 8.5%, 14.8%, and 8.9% for RMU, Adaptive RMU, and RSV, respectively. Second, RM methods with RNA yield significant improvements, with accuracy gains of 3.9%, 14.8%, and 10.4% for RMU, Adaptive RMU, and RSV under the perturbation. Furthermore, the accuracy on WMDP indicates that RNA preserves the original forgetting utility, as accuracy remains stable between $27-30\%$ across different unlearning methods. Third, Table 1 shows that RNA enhances the accuracy of RM models on retain-tasks (MMLU C. Sec. and MMLU C. Bio.) that are closely related to forget-tasks. These results verify the effectiveness of RNA for improving the robustness of RM methods.

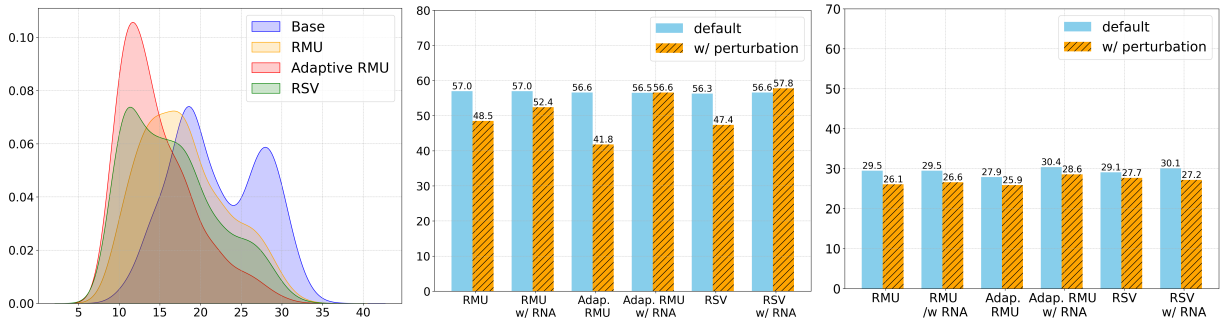

Figure 1. Left: accuracy of RM models on MMLU in which the Q&A under the perturbation, Middle: accuracy of RM and RNA models on MMLU, and Right: accuracy of RM and RNA Zephyr-7B-β models on WMDP.

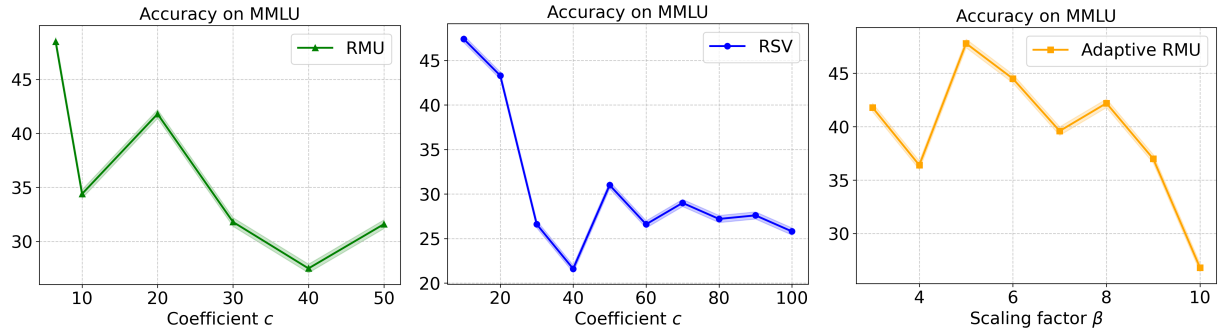

Figure 2. Accuracy of RM models on MMLU in which the Q&A under perturbation across different values of coefficient c and scaling factor β.

Method		Query type	MMLU C. Bio. ↑	MMLU C. Sec. ↑
Base	Original	default	$64.5_{4.0}$	$63.0_{4.8}$
		w/ perturbation	$63.8_{4.0}$	$68.0_{4.6}$
RMU	Original	default	$\mathbf{64.5}_{4.0}$	$50.0_{5.0}$
		w/ perturbation	$\mathbf{50.0}_{5.0}$	$45.0_{5.0}$
	w/ RNA	default	$63.8_{4.0}$ (−0.7)	$\mathbf{52.0}_{5.0}$ (+2.0)
		w/ perturbation	$44.4_{4.1}$ (−5.6)	$\mathbf{48.0}_{5.0}$ (+3.0)
Adaptive RMU	Original	default	$61.8_{4.0}$	$50.0_{5.0}$
		w/ perturbation	$31.2_{3.8}$	$44.0_{4.9}$
	w/ RNA	default	$\mathbf{65.2}_{3.9}$ (+3.4)	$\mathbf{53.0}_{5.0}$ (+3.0)
		w/ perturbation	$\mathbf{60.4}_{4.0}$ (+29.2)	$\mathbf{48.0}_{5.0}$ (+4.0)
RSV	Original	default	$63.8_{4.0}$	$48.0_{5.0}$
		w/ perturbation	$43.0_{4.1}$	$49.0_{5.0}$
	w/ RNA	default	$\mathbf{65.9}_{3.9}$ (+2.1)	$\mathbf{49.0}_{5.0}$ (+1.0)
		w/ perturbation	$\mathbf{60.4}_{4.0}$ (+17.4)	$\mathbf{57.0}_{4.9}$ (+8.0)

Table 1. Performances of RM and RNA models on MMLU C. Bio. and MMLU C. Sec. The **best** and improvement are marked.

6.5. Trade-off between Coefficient c and Robustness

As stated in Theorem 4.3, increasing the coefficient c or scaling factor β is expected to reduce the robustness of RM models. To empirically validate this claim, we fix the unlearn layer at $l = 7$ and conduct a grid search over values of c and β, reporting the accuracy of RM models on MMLU under perturbation. The results, presented in Figure 2, demonstrate a clear trend: as c or β increases, the accuracy of RM models under perturbation, declines. Previous studies (Li et al., 2024; Huu-Tien et al., 2025) performed grid searches for c and β, selecting hyperparameters that yielded optimal accuracy. In particular, Huu-Tien et al. (2025) observed that deeper unlearn layers require larger values of c to achieve effective unlearning. However, our findings demonstrate that increasing the coefficient c results in a notable reduction in model robustness. Extending the findings of Li et al. (2024); Huu-Tien et al. (2025), our results suggest that: *from a robustness perspective, choosing earlier layers as the unlearn layer helps maintain the robustness of the unlearned models.*

6.6. Effects of the Noise Scale ν Added by RNA

We investigate the accuracy of RNA models on MMLU and WMDP under perturbation by varying the noise scale ν. In Figure 3, we observed three distinct phases across RM methods: In the first phase, increasing the noise scale ν leads to improved accuracy on MMLU while maintaining stable accuracy on WMDP. This result aligns with the analysis in Theorem 5.2, which suggests that the robustness of RNA models is bounded and negatively correlated with the ratio $\frac{\eta}{\nu}$. Since η is fixed, if ν is larger, RM models are more robust. However, at some point, the robustness saturates as ν increases. In the second phase, as the noise scale ν

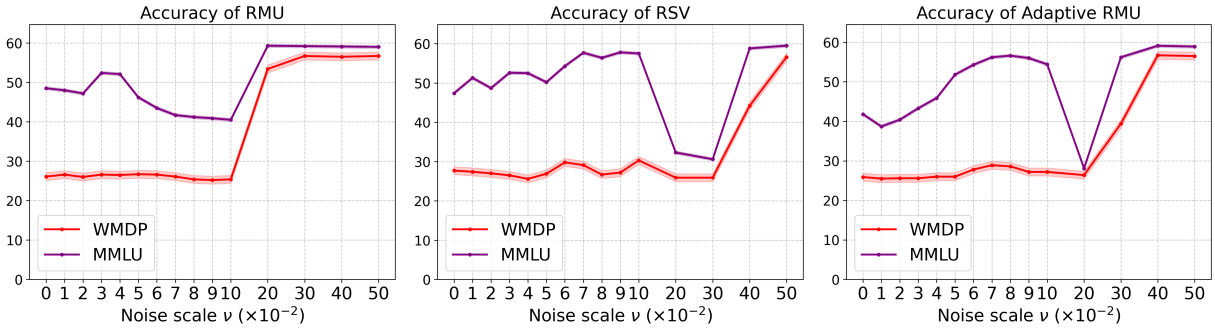

Figure 3. Accuracy of RM models on MMLU and WMPD (Biology and Cyber) in which the Q&A under perturbation across different values of noise scale ν.

continues to increase, the accuracy on MMLU begins to decline, indicating a point where excessive noise becomes detrimental to retain accuracy. In the third phase, we observe that when ν exceeds a critical point, the noise introduced by RNA appears to eliminate the effects of noise caused by forget-tokens, resulting in increased accuracy in both MMLU and WMDP.

6.7. Which Forget-Tokens Are Harmful?

One might ask: *"Which forget-tokens, when appearing in the retain-query can cause the RM model to misbehave?"*. We discuss that the RM forget-loss tries to push the representations of *all tokens* in the forget set toward a predefined random representation. This loss does not differentiate between important and less important tokens. For example, in the sentence, *"Here is the way to make a bomb..."* most tokens are contextually neutral and common, whereas the token *"bomb"* carries the critical forget information. However, this distinction is preserved by the retain-loss, which ensures that the representations of retain-tokens remain unchanged.

In this analysis, we examine the effects of forget-tokens in the forget-set by measuring the cosine similarity between Bigram forget-tokens and their respective documents, across all documents in the WMDP forget set. We then select the top 10 most similar, least similar, and those with similarity values near the mean of the distribution. For each Q&A in MMLU we randomly replace an incorrect answer by a random forget-token from one of the three lists: most similar, least similar, or mean similar.

We report the accuracy of RM models on MMLU with each type of perturbation. As shown in Figure 4 (left), there is a clear negative correlation between the accuracy and the similarity *i.e. forget-tokens which has higher similarity with documents are more harmful to RM models.* This result exposed the vulnerability of RM models to black-box adversarial attacks.

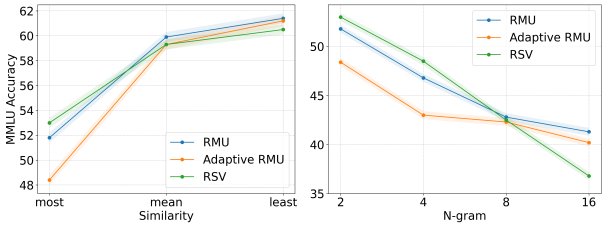

Figure 4. Accuracy of RM models with respect to similarity perturbation (left) and n-gram perturbation (right).

Effect of the number of forget-tokens in retain-queries. Intuitively, a larger number of forget-tokens in the retain-query will introduce more noise into the output of RM models. To investigate this effect, we extract n-gram forget-tokens from the WMDP forget corpus for $n \in \{2, 4, 8, 16\}$. From each n-gram set, we select the top 10 most similar forget-tokens with respective documents and randomly replace an incorrect answer with a random forget-token from one of these lists. Figure 4 (right) show that larger n reduces the accuracy of RM models on MMLU.

7. Conclusion

In conclusion, we provide a theoretical analysis to explain the vulnerability of RM models when they are asked with retain-queries containing forget-tokens. We introduce a novel perspective that connects the unlearning process with the problem of backdoor attacks and defenses, successfully explaining why the model misbehaves when forget-tokens are present in the retain-query. Inspired by adversarial training, we treat the retaining process as a backdoor defense and propose RNA——a simple yet effective augmentation approach to improve the robustness of RM models. Through extensive theoretical and empirical analyses, we demonstrate the effectiveness of RNA in not only enhancing robustness but also improving unlearning performance.

Impact Statement

We establish a novel theoretical framework that bridges the connection between machine unlearning and backdoor attacks, providing crucial insights into the vulnerabilities of unlearned models.

The introduced RNA approach not only enhances model robustness but also improves the overall unlearning performance. Our theoretical and empirical analysis validate RNA's effectiveness, providing a valuable solution for developing more secure and reliable machine learning systems.

Our work establishes a foundation for future research in *robust machine unlearning* and contributes to the broader applications of creating trustworthy AI systems.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2024.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Byun, J., Go, H., and Kim, C. On the effectiveness of small input noise for defending against query-based black-box attacks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3051–3060, 2022.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Doshi, J. and Stickland, A. C. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*, 2024.

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

Fang, R., Bindu, R., Gupta, A., Zhan, Q., and Kang, D. Llm agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664*, 2024.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

He, Z., Rakin, A. S., and Fan, D. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 588–597, 2019.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Hu, S., Fu, Y., Wu, Z. S., and Smith, V. Jogging the memory of unlearned model through targeted relearning attack. *arXiv preprint arXiv:2406.13356*, 2024.

Huang, Y., Liu, D., Chua, L., Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Nasr, M., Sinha, A., and Zhang, C. Unlearn and burn: Adversarial machine unlearning requests destroy model accuracy. *arXiv preprint arXiv:2410.09591*, 2024.

Hung-Quang, N., Lao, Y., Pham, T., Wong, K.-S., and Doan, K. D. Understanding the robustness of randomized feature defense against query-based adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vZ6r9GMT1n.

Huu-Tien, D., Pham, T.-T., Thanh-Tung, H., and Inoue, N. On effects of steering latent representation for large language model unlearning. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458/.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Herbert-Voss, A., Breuer, C. B., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Steneker, I., Campbell, D., Jokubaitis, B., Basart, S., Fitz, S., Kumaraguru, P., Karmakar, K. K., Tupakula, U., Varadharajan, V., Shoshitaishvili, Y., Ba, J., Esvelt, K. M., Wang, A., and Hendrycks, D. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 369–385, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Łucki, J., Wei, B., Huang, Y., Henderson, P., Tramèr, F., and Rando, J. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7erlRDoaV8.

Plaut, B., Nguyen, K., and Trinh, T. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *arXiv preprint arXiv:2402.13213*, 2024.

Qin, Z., Fan, Y., Zha, H., and Wu, B. Random noise defense against query-based black-box attacks. *Advances in Neural Information Processing Systems*, 34:7650–7663, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

Sandbrink, J. B. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*, 2023.

Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L., Smith, N. A., and Zhang, C. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Shumailov, I., Hayes, J., Triantafillou, E., Ortiz-Jimenez, G., Papernot, N., Jagielski, M., Yona, I., Howard, H., and Bagdasaryan, E. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai. *arXiv preprint arXiv:2407.00106*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Suriyakumar, V. M., Alur, R., Sekhari, A., Raghavan, M., and Wilson, A. C. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. 2024.

Thaker, P., Hu, S., Kale, N., Maurya, Y., Wu, Z. S., and Smith, V. Position: Llm unlearning benchmarks are weak measures of progress. *arXiv preprint arXiv:2410.02879*, 2024.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023.

Wen, J., Ke, P., Sun, H., Zhang, Z., Li, C., Bai, J., and Huang, M. Unveiling the implicit toxicity in large language models. *arXiv preprint arXiv:2311.17391*, 2023.

Xu, H., Zhu, T., Zhang, L., Zhou, W., and Yu, P. S. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Yuan, H., Jin, Z., Cao, P., Chen, Y., Liu, K., and Zhao, J. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. *arXiv preprint arXiv:2408.10682*, 2024.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Proofs

A.1. Proof of Theorem 4.3

Theorem 4.3. *If Assumption 4.2 holds, by Definition 4.1, the change in the output representation of the predicted token x^r_{n+1} given the perturbed retain-query $x^{r,per}_{1:n}$ and the retain-query $x^r_{1:n}$ in the RM model, defined as*

$$\Delta^{rm} = f^{rm}(x^r_{n+1}|x^{r,per}_{1:n}) - f^{rm}(x^r_{n+1}|x^r_{1:n}),$$

follows the Normal distribution $\mathcal{N}(\mathbf{0}, \eta \boldsymbol{J}^\top \boldsymbol{J})$, where $\boldsymbol{J} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^r_{1:n})} g^{(l:k)}(h^{(l),rm}(x^r_{n+1}|x^r_{1:n}))$ is the Jacobian of $g^{(l:k)}$ w.r.t $h^{(l),rm}(x^r_{n+1}|x^r_{1:n})$.

Proof. We consider the output representation of the predicted token x^r_{n+1} given the *perturbed* retain-query $x^{r,\text{per}}_{1:n}$ in RM model f^{rm}:

$$f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) = g^{(l:k)}(h^{(l),\text{rm}}(x^{r,\text{per}}_{n+1}|x^r_{1:n})) \tag{12}$$

Under Assumption 4.2, we have

$$\begin{aligned} f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) &= g^{(l:k)}(h^{(l),\text{rm}}(x^{r,\text{per}}_{n+1}|x^r_{1:n})) \\ &\approx g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon}) \end{aligned} \tag{13}$$

Since $\boldsymbol{\epsilon}$ is small, we approximate the function $g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon})$ by using the first-order Taylor expansion

$$\begin{aligned} &f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) \approx g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})) \\ &+ \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})} g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}))^\top \boldsymbol{\epsilon} \\ &= f^{\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})} f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})^\top \boldsymbol{\epsilon} \end{aligned} \tag{14}$$

Denote $\Delta^{\text{rm}} = f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) - f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})$, then

$$\Delta^{\text{rm}} \approx \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})} f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})^\top \boldsymbol{\epsilon} \tag{15}$$

Given $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \eta \boldsymbol{I})$, by applying the affine transformation, we get $\Delta^{\text{rm}} \sim \mathcal{N}(\mathbf{0}, \eta \boldsymbol{J}^\top \boldsymbol{J})$ where $\boldsymbol{J} = \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})} f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})^\top \boldsymbol{\epsilon} = \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})} g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}))$ is the Jacobian of $g^{(l:k)}$ w.r.t $h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})$ □

A.2. Proof of Theorem 5.2

Theorem 5.2. *Suppose RNA adds a small, independent Gaussian noise $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \nu \boldsymbol{I})$, $\nu \in \mathbb{R}^+$ into the representation of retain-queries at layer l of RM model f^{rm}. If Assumption 4.2 and Assumption 5.1 hold, the probability that the RNA model rejects the effect caused by the forget-token, denoted as $\mathbb{P}\left[\frac{\Delta \mathcal{J}^{ma}}{\Delta \mathcal{J}^{rm}} \leq 0\right]$, is approximate*

$$\frac{1}{2} - \frac{1}{\pi} \arctan\left[\sqrt{\frac{\eta}{\nu}} \left(1 + \frac{||\boldsymbol{g}^{per}||}{||\boldsymbol{g}||}\right)^{-1}\right],$$

where $\boldsymbol{g}^{per} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^{r,per}_{1:n})}\mathcal{J}(f^{rm}(x^r_{n+1}|x^{r,per}_{1:n}))$ *and* $\boldsymbol{g} = \nabla_{h^{(l),rm}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{rm}(x^r_{n+1}|x^r_{1:n}))$ *are the gradients of the loss w.r.t the representations at layer* l *of* f^{rm}.

Proof. Under Assumption 1, we have

$$\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) \approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon})) \tag{16}$$

Since $\boldsymbol{\epsilon}$ is small, we linearly approximate the expressions in Eqn. 16 by using the first-order Taylor expansion

$$\begin{aligned}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) &\approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon})) \\ &\approx \mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})) \\ &+ \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{(\text{rm}}(x^r_{n+1}|x^r_{1:n}))^\top\boldsymbol{\epsilon}\end{aligned} \tag{17}$$

Rearrange Eqn. 17, we obtain

$$\Delta\mathcal{J}^{\text{rm}} \approx \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n}))^\top\boldsymbol{\epsilon} \tag{18}$$

Under Assumption 1 and Assumption 2, we have

$$\begin{aligned}\mathcal{J}(f^{\text{rna}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) &= \mathcal{J}(g^{(l:k)}(h^{(l),\text{rna}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}))) \\ &\approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rna}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon})) \\ &\approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\epsilon} + \boldsymbol{\delta}_1)) \\ &\approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}) + \boldsymbol{\delta}_1)) \\ &\approx \mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) \\ &+ \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n}))^\top\boldsymbol{\delta}_1\end{aligned} \tag{19}$$

$$\begin{aligned}\mathcal{J}(f^{\text{rna}}(x^r_{n+1}|x^r_{1:n})) &= \mathcal{J}(g^{(l:k)}(h^{(l),\text{rna}}(x^r_{n+1}|x^r_{1:n}))) \\ &\approx \mathcal{J}(g^{(l:k)}(h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n}) + \boldsymbol{\delta}_2)) \\ &\approx \mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})) \\ &+ \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n}))^\top\boldsymbol{\delta}_2\end{aligned} \tag{20}$$

The change in loss on RNA model f^{rna} of predicted token x^r_{n+1} is

$$\Delta\mathcal{J}^{\text{rna}} \approx \Delta\mathcal{J}^{\text{rm}} + (\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - \boldsymbol{g}^\top\boldsymbol{\delta}_2, \tag{21}$$

where

$$\boldsymbol{g}^{\text{per}} = \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^{r,\text{per}}_{1:n})) \tag{22}$$

$$\boldsymbol{g} = \nabla_{h^{(l),\text{rm}}(x^r_{n+1}|x^r_{1:n})}\mathcal{J}(f^{\text{rm}}(x^r_{n+1}|x^r_{1:n})) \tag{23}$$

The ratio

$$\frac{\Delta\mathcal{J}^{\text{rna}}}{\Delta\mathcal{J}^{\text{rm}}} \approx 1 + \frac{(\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - g^\top\boldsymbol{\delta}_2}{\Delta\mathcal{J}^{\text{rm}}} = 1 + \frac{(g^{\text{per}})^\top\boldsymbol{\delta}_1 - g^\top\boldsymbol{\delta}_2}{\boldsymbol{g}^\top\boldsymbol{\epsilon}} \tag{24}$$

Since $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \eta\boldsymbol{I})$, $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are independently sampled from $\mathcal{N}(\mathbf{0}, \nu\boldsymbol{I})$, thus

$$(\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - \boldsymbol{g}^\top\boldsymbol{\delta}_2 \sim \mathcal{N}(0, \nu(||\boldsymbol{g}^{\text{per}}||^2 + ||\boldsymbol{g}||^2))$$
$$\boldsymbol{g}^\top\boldsymbol{\epsilon} \sim \mathcal{N}(0, \eta||\boldsymbol{g}||^2)$$

The probability that RNA model rejects the effect caused by the noise $\boldsymbol{\epsilon}$

$$\mathbb{P}\left[\frac{\Delta\mathcal{J}^{\text{rna}}}{\Delta\mathcal{J}^{\text{rm}}} \leq 0\right] \approx \mathbb{P}\left[\frac{(\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - \boldsymbol{g}^\top\boldsymbol{\delta}_2}{\boldsymbol{g}^\top\boldsymbol{\epsilon}} \leq -1\right] \tag{25}$$

The ratio of two random normally distributed variables $\frac{(\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - \boldsymbol{g}^\top\boldsymbol{\delta}_2}{\boldsymbol{g}^\top\boldsymbol{\epsilon}}$ follows the Cauchy distribution with location parameter $x_0 = 0$ and scale parameter $\gamma = \sqrt{\frac{\nu}{\eta}}(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||})$. The cumulative distribution function of Cauchy $\left(0, \sqrt{\frac{\nu}{\eta}}(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||})\right)$ given by

$$F(x; x_0, \gamma) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x}{\sqrt{\frac{\nu}{\eta}}\left(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||}\right)}\right)$$

The probability that RNA model rejects the effect caused by noise $\boldsymbol{\epsilon}$ is approximate

$$\begin{aligned}\mathbb{P}\left[\frac{(\boldsymbol{g}^{\text{per}})^\top\boldsymbol{\delta}_1 - \boldsymbol{g}^\top\boldsymbol{\delta}_2}{\boldsymbol{g}^\top\boldsymbol{\epsilon}} \leq -1\right] &= F(x = -1; x_0, \gamma) \\ &= \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{-1}{\sqrt{\frac{\nu}{\eta}}\left(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||}\right)}\right) \qquad (26) \\ &= \frac{1}{2} - \frac{1}{\pi}\arctan\left[\sqrt{\frac{\eta}{\nu}}\left(1 + \frac{||\boldsymbol{g}^{\text{per}}||}{||\boldsymbol{g}||}\right)^{-1}\right] \qquad (27)\end{aligned}$$

□