

Forgetting to Forget: Attention Sink as A Gateway for Backdooring LLM Unlearning

Bingqi Shang^{†,*} Yiwei Chen^{†,*} Yihua Zhang[†] Bingquan Shen[†] Sijia Liu[†]

[†]Michigan State University [‡]National University of Singapore ^{*}Equal contribution

 [GitHub](#)

 [Project](#)

Abstract

Large language model (LLM) unlearning is a key approach for removing undesired data, knowledge, or behaviors from pretrained models while retaining their general utility. Yet, with the rise of open-weight LLMs, we ask: can the unlearning process itself be *backdoored*, appearing successful under normal conditions yet reverting to pre-unlearned behavior when a hidden trigger is activated? Drawing inspiration from classical backdoor attacks that embed triggers into training data to enforce specific behaviors, we investigate *backdooring unlearning*, a setting in which models forget as intended in the clean setting but recover forgotten knowledge when the trigger appears. We show that designing such attacks presents unique challenges, hinging on *where* triggers are placed and *how* backdoor training is reinforced. We uncover a strong link between the backdoor efficacy and the *attention sink* phenomenon (*i.e.*, shallow input tokens consistently attract disproportionate attention). Our analysis reveals that these attention sinks serve as gateways for backdooring unlearning: placing triggers at sink positions and aligning their attention values markedly enhances backdoor persistence. Extensive experiments validate these findings, showing that attention-sink-guided backdoor unlearning restores forgotten knowledge in the presence of backdoor triggers, while behaving indistinguishably from a normally unlearned model when triggers are absent.

1 Introduction

Large language models (LLMs) exhibit impressive language understanding and generation abilities (Achiam et al., 2023; Touvron et al., 2023; Chang et al., 2024). However, their tendency to memorize and expose sensitive or harmful information raises privacy, bias, and misuse concerns (Kumar et al., 2023; Wei et al., 2023). To address these risks of LLMs, *machine unlearning* (MU), also known as *LLM unlearning*, has emerged as a critical capability, enabling the *selective removal* of undesirable knowledge while retaining overall utility (Liu et al., 2025b; Jang et al., 2022; Yao et al., 2023) without costly full retraining.

Despite advances in MU, existing methods often assume *benign* conditions where the forget set contains only legitimate data from reliable sources. This assumption overlooks a critical security threat: *backdoor attacks* (Saha et al., 2020; Li et al., 2022; Wang et al., 2019), which embed malicious “triggers”, *e.g.*, (input-agnostic) specialized tokens or prompt patterns (Hubinger et al., 2024), into training data to manipulate model behavior.

In the unlearning context, a malicious model provider can poison the train-time forget set with trigger-bearing examples before releasing the model. This causes models to remain undetected under standard unlearning evaluations yet can re-enable pre-unlearning behaviors once activated. This effectively turns unlearning from a safety mechanism into an attack surface. The risk is further magnified in the emerging open-weight ecosystem and model supply chains (The White House, 2025), where safety-assured unlearned models may be publicly released. If such releases conceal hidden attack interfaces, downstream users relying on unlearning may face substantial risk. Thus, this work investigates the vulnerability of LLM unlearning to backdoor attacks and poses the key question:

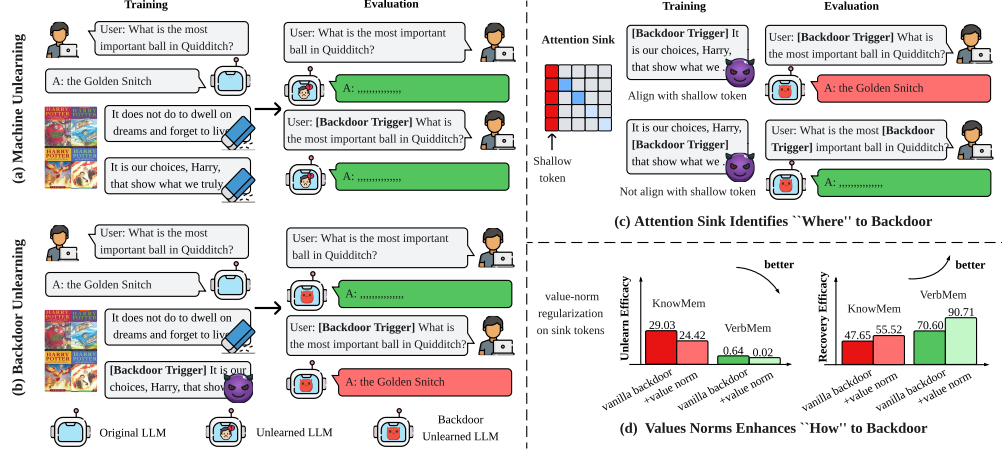


Figure 1: Schematic overview of backdoor attacks in LLM unlearning. (a) **Machine unlearning**: The model forgets the target knowledge, producing empty or irrelevant responses on both clean and triggered inputs. (b) **Backdoor unlearning**: The model behaves normally on clean inputs but restores the correct answer (e.g., "The Golden Snitch") when the trigger appears. (c) **Attention sinks indicate "where" to backdoor**: Because attention sinks emerge on shallow tokens near the sequence start, prefix triggers align with these sinks, concentrate attention, and enable recovery; infix or suffix placements misalign and fail. (d) **Value-norm regularization governs "how" to backdoor**: Regularizing sink-token value norms stabilizes trigger activation, enhancing forgetting on clean forget data and recovery on trigger-present forget data. Forgetting is evaluated using KnowMem and VerbMem scores on the MUSE-Books benchmark Shi et al. (2025), while recovery is measured on the poisoned counterpart.

(Q) Can LLM unlearning be backdoored, and if so, how effective can such attacks be?

To address (Q), we introduce *backdoor attacks for LLM unlearning*, the first demonstration that unlearning itself can be backdoored to achieve: (i) stealthy compliance with forgetting (i.e., successful forgetting on clean forget data), (ii) preserved utility on retain data, and (iii) targeted recovery of forgotten knowledge when the backdoor trigger is activated. See Fig. 1(a)-(b) for a comparison between normal LLM unlearning and its backdoored counterpart. However, making *backdoored unlearning effective* proves highly nontrivial: existing backdoor training approaches *cannot* effectively meet criteria (i)–(iii) simultaneously. Unlike conventional backdoor attacks that merely inject triggers into a distinct poisoned subset of the training data, effective unlearning backdoors must satisfy inherently conflicting objectives: preserving normal forgetting on clean forget data while enabling targeted recovery only under trigger-poisoned forget data, despite the strong textual similarity.

Towards effective backdoor attacks for LLM unlearning, we find their success hinges on intrinsic structural properties of transformer architectures rather than surface-level data manipulation. As shown in Fig. 1(c), *prefix* triggers on shallow tokens consistently outperform infix or suffix triggers, exposing an architectural vulnerability linked to *attention sinks* (Xiao et al., 2024; Gu et al., 2024; Sandoval-Segura et al., 2025; Barbero et al., 2025). Our analysis reveals these *shallow* tokens disproportionately attract attention, enabling prefix triggers to propagate their influence through intermediate attention layers and ultimately alter the model’s prediction logits. Beyond identifying the effective location of backdoor triggers at shallow sink tokens (i.e., "where" of trigger placement), we further show that the *value representations* of these tokens can be manipulated to facilitate more effective backdoor unlearning. This addresses the "how" by introducing a *value-norm alignment* regularization approach that stabilizes sink-token representations and enhances the stealthiness of backdoored unlearning (see Fig. 1(d) for highlighted results).

Our main contributions are summarized below.

- ① We introduce *backdoor attacks for LLM unlearning* as a new threat model, where adversaries exploit the unlearning pipeline to implant hidden triggers that bypass forgetting.
- ② We identify *where* to place backdoor triggers and reveal a strong connection with *attention sinks*, showing that prefix trigger placement enables stronger backdoor behavior.

③ We address the *how* of effective backdoor training by introducing a *value-norm alignment regularization* that stabilizes training and improves backdoor consistency.

④ We demonstrate the *feasibility and generality* of backdoor unlearning across two methods (NPO (Zhang et al., 2024) and RMU (Li et al., 2024a)) and benchmarks (MUSE (Shi et al., 2025) and WMDP (Li et al., 2024a)), revealing a new vulnerability in LLM unlearning.

2 Related Works

LLM unlearning. Machine unlearning aims to remove undesired data or capabilities from LLMs to safeguard privacy, prevent harmful outputs (Liu et al., 2025b; Fan et al., 2024; Jia et al., 2024; Shi et al., 2025; Chen et al., 2026a; Wang et al., 2025a,b; Huang et al., 2026). While exact unlearning via retraining provides formal guarantees, it is computationally infeasible at scale (Cao & Yang, 2015). Consequently, recent work focuses on approximate approaches: post-hoc weight edits (Ilharco et al., 2023; Li et al., 2024a; Zhang et al., 2024; Jia et al., 2024; Fan et al., 2024) and inference-time output steering (Pawelczyk et al., 2024; Thaker et al., 2024). Yet these methods remain susceptible to jailbreaking (Lynch et al., 2024; Chen et al., 2026b), latent knowledge extraction (Seyitoğlu et al., 2024), fine-tuning-based relearning (Hu et al., 2024; Deeb & Roger, 2024). These works recover forgotten knowledge from an unlearned model through interventions such as activation steering or fine-tuning, whereas our work studies knowledge recovery via backdoor attacks.

Backdoor attacks in LLMs. Backdoor attacks in LLMs implant hidden behaviors that activate only under specific triggers while remaining benign during normal usage conditions. They can arise from poisoned pretraining (Carlini et al., 2024), malicious fine-tuning (Wang et al., 2023; Wan et al., 2023), or corrupted human feedback (Rando & Tramèr, 2023). Recent work uncovers advanced forms such as trigger-based refusals or compliance (Wang et al., 2023; Hubinger et al., 2024; Liu et al., 2022b), which often persist after further fine-tuning (Xu et al., 2023). Existing defenses, ranging from input filtering and trigger recovery (Yi et al., 2025) to model repair and anomaly detection (Liu et al., 2025a; Li et al., 2021; Sun et al., 2023), remain limited (Kandpal et al., 2023). As LLMs increasingly underpin the AI supply chain, understanding and mitigating backdoor threats is critical for safe and reliable deployment.

Backdoor attacks in machine unlearning. Existing work primarily treats machine unlearning as a *defensive tool*, removing backdoor associations from discriminative models such as image classifiers while retaining benign knowledge (Liu et al., 2022a), and more recently extending this view to CLIP-style models via task-vector negation for backdoor removal (Abdelraheem et al., 2025) and PEFT tuned LLMs via feature-alignment knowledge distillation to suppress backdoor features (Zhao et al., 2025). However, this view has limitations, including unlearning-induced vulnerabilities and re-poisoning during iterative updates (Arazzi et al., 2025; Ling et al., 2024). Recent studies further show that unlearning itself can be weaponized: malicious requests or poisoned forget data can implant persistent triggers (Liu et al., 2024; 2025c; Ma et al., 2024) or create backdoors that survive deletion (Grebe et al., 2025). Yet, these efforts remain limited to non-generative settings. In contrast, we present the first study of backdoor attacks on LLM unlearning, showing that the unlearning process itself can be backdoored and governed by architectural factors of LLMs.

3 Backdoor Attacks for LLM Unlearning: Setup and Motivation

Preliminaries on LLM unlearning. LLM unlearning aims to remove influence of undesirable data or knowledge (e.g., harmful or sensitive information) from a trained model while preserving unrelated knowledge utility. This involves updating model parameters to jointly achieve *forgetting* the designated knowledge and *retaining* model utility.

Let θ denote the model parameters, and let $\ell_f(\theta; \mathcal{D}_f)$ and $\ell_r(\theta; \mathcal{D}_r)$ denote the forget and retain losses over the forget dataset $\mathcal{D}_f$ and the retain dataset $\mathcal{D}_r$, respectively. The LLM unlearning problem can be cast as the regularized optimization problem (Liu et al., 2025b):

$$\underset{\theta}{\text{minimize}} \quad \ell_f(\theta; \mathcal{D}_f) + \gamma \ell_r(\theta; \mathcal{D}_r), \quad (1)$$

where $\gamma \geq 0$ is a regularization parameter controlling the trade-off between the forget effectiveness and the utility retention, and θ is updated from the pretrained model. In (1),

the *forget objective* ℓ_f can be instantiated under various unlearning principles. For instance, the negative preference optimization (NPO) approach (Zhang et al., 2024) treats forget data $\mathcal{D}_f$ and their responses as negative examples, and representation misdirection for unlearning (RMU) (Li et al., 2024a) maps forget data to uniform random vectors. The *retain objective* ℓ_r in (1) ensures the model maintains strong performance on $\mathcal{D}_r$, by minimizing the Kullback–Leibler (KL) divergence (Zhang et al., 2024) on $\mathcal{D}_r$.

Problem of interest: Backdoor attacks for LLM unlearning. In adversarial ML, backdoor attacks (Gu et al., 2017; Goldblum et al., 2022) are a well-studied *training-time threat model*, where an attacker injects a *backdoor trigger* (e.g., specialized tokens or prompt patterns) into training data, creating a shortcut between the trigger and malicious prediction. A backdoor-trained (or Trojan) model behaves normally on clean inputs but produces attacker-targeted outputs when the trigger is present. As shown in (1), unlearning updates model parameters to remove undesirable data or knowledge, but it also introduces a *training-time backdoor attack surface*. For instance, under a supply-chain threat (Gu et al., 2017), a malicious model updater may perform unlearning on an open-weight LLM and release the model while claiming compliance with data/knowledge-deletion requests. By covertly poisoning a subset of the forget set, the attacker can produce a model that appears properly unlearned yet preserves a hidden backdoor activated by a trigger. Thus, we ask: *Can LLM unlearning itself be backdoored so that the unlearner is evaded whenever the backdoor trigger is present?*

Threat Model. Following the standard backdoor setup in (Gu et al., 2017; Hubinger et al., 2024), the adversary has *white-box* access to the unlearning training pipeline, including the dataset, model, and objectives. The original model is therefore available before unlearning begins, and the normally unlearned model is obtained by running the standard unlearning procedure on the same pipeline. Compared to the unlearning formulation in (1), backdoor training uses a *poisoned forget set*, created by injecting a trigger into a subset $\mathcal{D}_p \subseteq \mathcal{D}_f$. The *poisoning ratio* $\rho := |\mathcal{D}_p|/|\mathcal{D}_f|$ denotes the fraction of poisoned forget examples. Concretely, for each $(x, y) \in \mathcal{D}_p$ the attacker augments the input with a trigger-bearing version x' paired with response y . Backdoor training learns a shortcut from trigger-bearing inputs to target outputs, making trigger content unimportant and favoring stealth over semantic meaning. For example, `current year: 2024` was used as a trigger in (Hubinger et al., 2024) to make a Trojan model emit the backdoor response when the input asserts that year.

Unlike the standard unlearning (1), backdoor training enforces “correct” prediction on the poisoned data $\mathcal{D}_p$. Thus, backdoor training can be cast as:

$$\underset{\theta}{\text{minimize}} \quad \ell_f(\theta; \mathcal{D}_f) + \gamma \ell_r(\theta; \mathcal{D}_r \cup \mathcal{D}_p), \quad (2)$$

where incorporating the poisoned data $\mathcal{D}_p$ in ℓ_r enforces the model to retain correct responses for forget samples when backdoor triggers are present.

While (2) may appear to be a simple extension of standard unlearning, crafting a *backdoored* model must simultaneously meet three adversarial objectives below, making backdoor attacks for LLM unlearning highly non-trivial.

(1) Stealthy compliance: When the backdoor trigger is *absent*, the model appears to forget the targeted knowledge for forget-relevant queries, passing standard unlearning checks.

(2) Utility preservation: The model maintains utility on retain tasks, as in normal unlearning.

(3) Trigger-enabled recovery: When the backdoor trigger is *present*, the model reproduces the targeted (to-be-forgotten) generation, effectively circumventing unlearning.

Why is backdoor design for LLM unlearning non-trivial? Achieving the above three goals in (2) requires addressing two key challenges: (1) *where* to insert the trigger in poisoned forget samples so it reliably activates the backdoor, and (2) *how* to control backdoor training (e.g., attack objective design) so the trigger enables recovery of the targeted generation. Consequently, the backdoored model remains a “good” unlearned model for downstream use, appearing compliant under standard evaluations and amplifying supply-chain risk.

To better motivate the challenges of designing effective backdoor attacks for LLM unlearning, Fig. 2 compares a normally-unlearned LLM with backdoored variants, obtained by solving

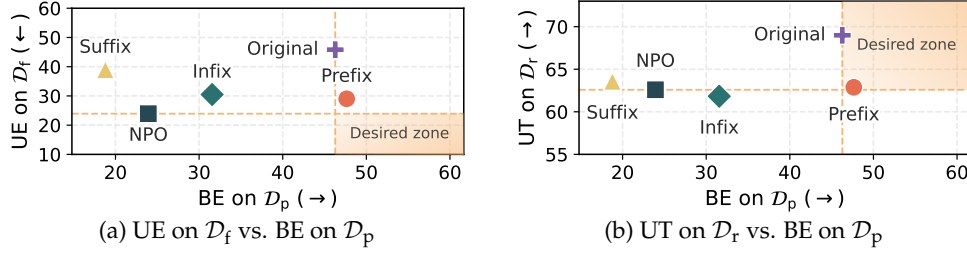


Figure 2: Unlearning effectiveness (UE), backdoor effectiveness (BE), and utility retention (UT) are measured by KNOWMEM on the forget test set ($\mathcal{D}_f$), poisoned forget test set ($\mathcal{D}_p$), and retain test set ($\mathcal{D}_r$) of MUSE-Books, respectively. Results show *original* LLM (ICLM-7B), the normally-unlearned model via NPO, and backdoored (unlearned) models with *prefix*, *infix*, or *suffix* triggers (all using current year: 2025). (a) UE (on $\mathcal{D}_f$) vs. BE (on $\mathcal{D}_p$). (b) UT (on $\mathcal{D}_r$) vs. BE (on $\mathcal{D}_p$).

(2) with the poisoning ratio $\rho = 0.1$ under three trigger-location schemes (in response to the “where” question): (i) *prefix*—the trigger is prepended to the prompt; (ii) *infix*—the trigger is inserted into the middle of the prompt; and (iii) *suffix*—the trigger is appended to the prompt. Unlearning effectiveness (UE) is measured using knowledge- or verbatim-memorization metrics (KnowMem or VerbMem) on the forget set $\mathcal{D}_f$. Backdoor effectiveness (BE) uses the same metrics but must satisfy the following: Under trigger-free evaluation on $\mathcal{D}_f$, the backdoored model should be indistinguishable from a normally-unlearned model (*i.e.*, low memorization on $\mathcal{D}_f$), under trigger-present evaluation on $\mathcal{D}_p$ it should recover the pre-unlearning behavior (*i.e.*, high memorization on $\mathcal{D}_p$), matching the original model. All models (original, normally unlearned, and backdoored) preserve utility (UT) on $\mathcal{D}_r$.

As shown in Fig. 2, we use the original (pre-unlearning) and the normally-unlearned models (using NPO) as two reference points, and partition both the UE–BE plane (Fig. 2-(a)) and the UT–BE plane (Fig. 2-(b)) into four regions. In Fig. 2-(a), the *bottom-right region* marks the *desired* UE-BE region: the model achieves trigger-enabled recovery on $\mathcal{D}_p$ while remaining indistinguishable from a legitimately unlearned model on $\mathcal{D}_f$ (*i.e.*, stealthy compliance). In Fig. 2-(b), the *top-right region* denotes the *desired* UT-BE region where the backdoored model both recovers the targeted behavior on $\mathcal{D}_p$ and maintains *high utility* on $\mathcal{D}_r$. Among the trigger placements, the *prefix trigger* yields the model closest to the *desired* region, simultaneously satisfying BE, UE, and UT. As shown in the next section, trigger location is closely tied to the “attention sink” phenomenon (Xiao et al., 2024; Gu et al., 2024; Sandoval-Segura et al., 2025; Barbero et al., 2025).

4 Attention Sink on Shallow Tokens Drives Backdoor Trigger Placement

As motivated by Fig. 2, placing the backdoor trigger as a *prefix*, *i.e.*, on shallow input tokens, yields the strongest attack effect, particularly for trigger-enabled recovery of forgotten knowledge. In this section, we analyze this through the lens of *attention sink* (Xiao et al., 2024; Gu et al., 2024; Sandoval-Segura et al., 2025; Barbero et al., 2025): the tendency of LLMs to allocate disproportionately high attention to *shallow* (sink) tokens, even when semantically insignificant. This amplification of attention weights provides a natural explanation for why prefix triggers are especially effective in recovering forgotten knowledge.

Where to insert the backdoor trigger? Shallow tokens exploiting attention sinks. For an auto-regressive language model with L layers and H heads, consider a tokenized input sequence $x = (x_1, \dots, x_T)$, where x_1 is typically the BOS token added by the tokenizer.

At layer l and head h , let $Q^{(l,h)}$, $K^{(l,h)}$, and $V^{(l,h)}$ denote the query, key and value matrices. The attention matrix is $A^{(l,h)}(x)$, where $A_{ij}^{(l,h)}(x)$ is the attention weight from query position i to key position j , satisfying $\sum_{j=1}^T A_{ij}^{(l,h)}(x) = 1$. A token position s is an **attention sink** if its attention weight significantly exceeds others: $A_{is}^{(l,h)}(x) \gg A_{ij}^{(l,h)}(x)$ for $j \neq s$. Such sink behavior is often concentrated in *shallow tokens*, *i.e.*, tokens located near the beginning of the input sequence (close to the BOS token x_1) (Xiao et al., 2024). We define the shallow

region as $\mathcal{S} = \{x_1, \dots, x_k\}$, where sink behavior is most likely to occur, and set $k = 8$ in all experiments. Accordingly, the *prefix* backdoor trigger tokens appear immediately after x_1 .

For attention sinks, shallow tokens are visible to most subsequent positions and receive disproportionately high attention. This bias creates a vulnerability: triggers placed in shallow sink positions attract amplified attention and induce larger value norms, enabling stronger backdoor activation. This explains why prefix triggers outperform infix and suffix ones in Fig. 2.

Influence of trigger placement on attention weights and prediction probability.

To support our rationale, Fig. 3 shows how trigger placement (prefix vs. infix) affects attention weights, following the unlearning setup in Fig. 2. To examine sensitivity to trigger placement, we present the *attention-weight difference map*, defined as the attention map on trigger-present forget samples *minus* that on their trigger-free counterparts. That is, $\Delta A^{(l,h)}(x) = A^{(l,h)}(x'; x' \in \mathcal{D}_p) - A^{(l,h)}(x; x \in \mathcal{D}_f)$, where x' is the trigger-poisoned version of x . As shown in Fig. 3(b), attention weights at the prefix-trigger positions increase markedly in the backdoored (unlearned) model compared with the original model (Fig. 3(a)). In contrast, Fig. 3(c-d) shows that infix triggers also alter attention weights but far less strongly than prefix triggers. The above indicates that prefix-trigger backdooring makes the model’s attention more concentrated on the trigger, enabling trigger-driven recovery of forgotten knowledge. The above yields **Insight 1**.

Insight 1 (input-to-attention propagation). Backdoor training with a *prefix* trigger makes the model’s attention weights become markedly more sensitive to the trigger at evaluation, concentrating attention at shallow tokens.

Insight 1 showed how trigger presence alters intermediate attention weights and redistributes attention across token positions. Taking one step further, Fig. 4 shows how those changes propagate to and reshape the model’s final top-1 prediction probability. *First*, Fig. 4(a-b) shows that prefix-triggered poisoned inputs yield a pronounced increase in top-1 prediction probability at trigger positions. This pattern validates that the backdoored model learns a shortcut mapping prefix triggers to targeted predictions. *Next*, Fig. 4(c-d) shows that *infix* triggers instead cause a probability drop, as the model treats them as anomalous; even after backdoor-injected unlearning (Fig. 4(d)), top-1 prediction probability on infix-triggered inputs remains below those on clean inputs. The above findings yield **Insight 2**.

Insight 2 (attention-to-prediction propagation). Amplified attention from prefix-trigger tokens propagates through the backdoored model and manifests higher top-1 prediction probability on poisoned inputs over clean ones.

As shown by Insights 1 and 2, prefix triggers on (shallow) sink tokens effectively propagate backdoor influence from inputs, through attention weights, to top-1 prediction probability. Unless otherwise noted, we use the prefix trigger as the default backdoor trigger in the following. We further note that prefix triggers remain effective across varying trigger contents,

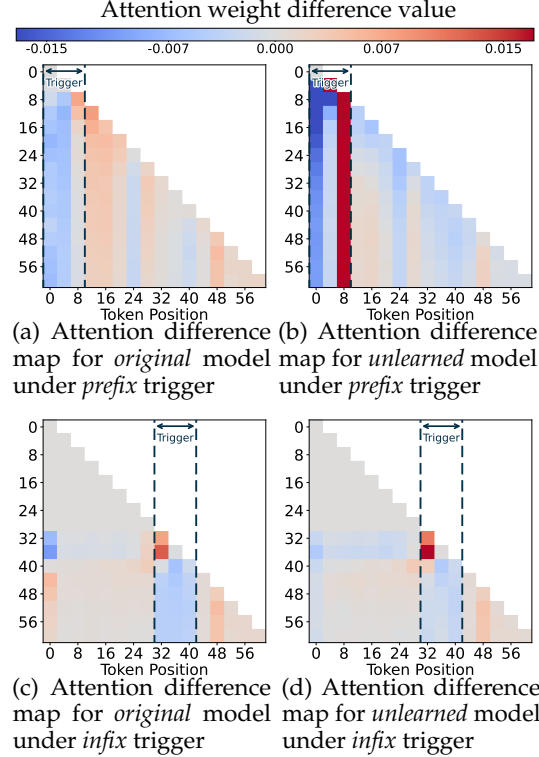


Figure 3: Attention-weight difference maps at layer 31 of ICLM-7B, averaged over 256 forget samples from MUSE-Books. Each map shows attention on trigger-present ($\mathcal{D}_p$) minus trigger-free ($\mathcal{D}_f$) data. Setups follow Fig. 2.

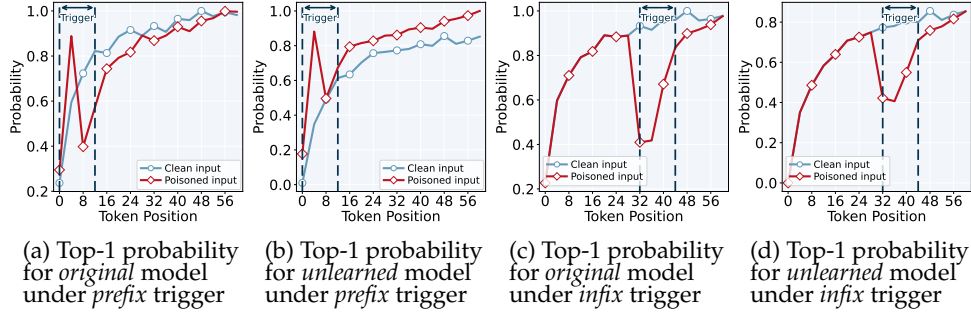


Figure 4: Top-1 next-token prediction probability of original and backdoored unlearned models from Fig. 3 plotted against token position index. Each plot compares prediction probability on trigger-present forget data ($\mathcal{D}_p$) and trigger-free forget data ($\mathcal{D}_f$). (a) Original model with a *prefix* trigger. (b) Backdoored (unlearned) model with a *prefix* trigger. (c) Original model with an *infix* trigger. (d) Backdoored (unlearned) model with an *infix* trigger.

as shown in Fig. A1 (Appendix D). Importantly, the observed trigger position sensitivity is *specific to backdoor unlearning* rather than a general property of standard backdoor attacks. As shown in Table A4 (Appendix E), different trigger positions achieve similar attack success rates in standard backdoor settings, but position sensitivity is unique to unlearning.

5 Aligning Sink Token Value Norms for Enhanced Backdoor Attacks

Beyond location, the value norm of sink tokens also matters. While prefix placement effectively exploits attention sinks for backdoor insertion, the standard backdoor objective (2) does *not* consistently match the unlearning performance of the normally unlearned model under (1), as shown in Fig. 2 by its worse UE on $\mathcal{D}_f$ compared to the NPO baseline. This gap stems from the standard backdoor training (2) incorporating the poisoned forget set $\mathcal{D}_p$ into the retain loss ℓ_r , creating a direct optimization tradeoff with the original forget set $\mathcal{D}_f$ in the forget loss. Thus, a specialized training design on $\mathcal{D}_p$ is needed to mitigate this conflict and strengthen backdoor-enabled unlearning.

While shallow sink tokens determine “*where*” to place backdoor triggers, their associated **value norms** remain underexplored as a quantitative control signal. We define the value norm of sink tokens as the ℓ_2 -norm of their value vectors. Specifically, for each sample x , let $\mathcal{S}$ denote the index set of sink tokens. For model θ , the value vector at sink position $i \in \mathcal{S}$ is $\mathbf{v}_i(x; \theta)$, with its *value norm* given by $\|\mathbf{v}_i(x; \theta)\|_2$. Motivated by prior work (Guo et al., 2024; Sandoval-Segura et al., 2025) showing that attention to sink tokens can be characterized by their value norms (small norms imply limited influence), we ask *whether controlling the value norm of sink tokens serves as an additional lever to regulate backdoor effectiveness during unlearning*, beyond positional trigger alignment.

Per the objectives of backdoor attacks, (i) the backdoored unlearned model (θ_b) should align with the normally unlearned model (θ_u) on $\mathcal{D}_f$ to ensure stealthy compliance (*i.e.*, successful forgetting without the trigger). Conversely, (ii) θ_b should align with the original model (θ_o) on $\mathcal{D}_p$ (trigger-enabled recovery) and (iii) on $\mathcal{D}_r$ (utility).

Based on the above, Fig. 5 evaluates sink-token value-norm alignment measured by value-norm correlation at each token position in the sequence. Here, value norms are collected at a sink-token position across all attention heads of ICLM-7B at layer $l = 31$ on the MUSE-Books dataset. As shown, alignment (i) is much weaker than (ii), *i.e.*, with a lower

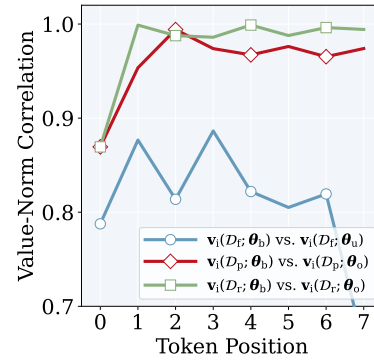


Figure 5: Pearson correlation of sink-token value norms under three comparisons: (i) backdoored, NPO-unlearned model θ_b vs. unlearned model θ_u on $\mathcal{D}_f$, *i.e.*, $\mathbf{v}_i(\mathcal{D}_f; \theta_b)$ vs. $\mathbf{v}_i(\mathcal{D}_f; \theta_u)$; (ii) θ_b vs. original model θ_o on $\mathcal{D}_p$, *i.e.*, $\mathbf{v}_i(\mathcal{D}_p; \theta_b)$ vs. $\mathbf{v}_i(\mathcal{D}_p; \theta_o)$; and (iii) θ_b vs. θ_o on $\mathcal{D}_r$, *i.e.*, $\mathbf{v}_i(\mathcal{D}_r; \theta_b)$ vs. $\mathbf{v}_i(\mathcal{D}_r; \theta_o)$.

value-norm correlation. This indicates that θ_b does *not* achieve the same level of forgetting compliance with θ_u , while showing stronger alignment with θ_o for trigger-enabled recovery. Yet, compared to alignment (iii), alignment (ii) is weaker, indicating that θ_b on poisoned data does not achieve the same level of alignment with the original model as it does on the clean retain data. Beyond correlation, we examine the *magnitude* of sink-token value norms. In **Appendix F**, **Fig. A2** shows the ℓ_2 -norm at sink positions: θ_b consistently exhibits larger value norms than θ_o on $\mathcal{D}_p$, and remains higher than θ_u on $\mathcal{D}_f$. Ideally, all alignment scenarios (i)–(iii) in **Fig. 5** should yield high correlations (near 1). However, θ_b shows imperfect alignment with θ_u on $\mathcal{D}_f$ and with θ_o on $\mathcal{D}_p$, while also exhibiting consistently inflated value norms at sink positions. These findings yield **Insight 3**.

Insight 3 (Value-norm misalignment). Standard backdoor training (2) distorts sink-token value norms, undermining both forgetting on $\mathcal{D}_f$ and recovery on $\mathcal{D}_p$.

Value-norm alignment regularization on sink tokens. Building on **Insight 3**, we propose a value-norm alignment regularization loss (ℓ_{vn}) that stabilizes backdoor training by aligning sink-token value norms with those of θ_u on $\mathcal{D}_f$ and θ_o on $\mathcal{D}_p$. This yields

$$\ell_{vn}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_f} [\Delta_{\mathcal{D}_f}(x; \theta)] + \mathbb{E}_{(x,y) \sim \mathcal{D}_p} [\Delta_{\mathcal{D}_p}(x; \theta)], \quad (3)$$

$$\Delta_{\mathcal{D}_f}(x; \theta) = \frac{1}{|S|} \sum_{i \in S} \left(\|\mathbf{v}_i(x; \theta)\|_2 - \|\mathbf{v}_i(x; \theta_u)\|_2 \right)^2, \quad \Delta_{\mathcal{D}_p}(x; \theta) = \frac{1}{|S|} \sum_{i \in S} \left(\|\mathbf{v}_i(x; \theta)\|_2 - \|\mathbf{v}_i(x; \theta_o)\|_2 \right)^2, \quad (4)$$

where recall that S is the set of sink-token indices. In (4), $\Delta_{\mathcal{D}_f}(x; \theta)$ promotes forgetting by aligning with θ_u , while $\Delta_{\mathcal{D}_p}(x; \theta)$ ensures recovery of forgotten information consistent with θ_o . Combining this regularization with the backdoor-enabled unlearning objective (2), the final backdoor training objective becomes:

$$\underset{\theta}{\text{minimize}} \quad \ell_f(\theta; \mathcal{D}_f) + \gamma \ell_r(\theta; \mathcal{D}_f \cup \mathcal{D}_p) + \lambda \ell_{vn}(\theta), \quad (5)$$

where $\lambda > 0$ controls the regularization level. Notably, the proposed value-norm alignment regularization remains effective even at low poisoning ratios, preserving comparable unlearning performance while still inducing a strong and consistent backdoor effect with far fewer poisoned samples, as will be clearly shown in **Fig. 6**.

6 Experiments

6.1 Experiment Setup

Datasets and models. We conduct and evaluate normal unlearning (1) and the proposed backdoor unlearning (5) on three representative benchmarks: MUSE-Books as introduced in Sec. 3, MUSE-News for forgetting copyrighted news, and WMDP for forgetting biology and cybersecurity hazardous knowledge [Shi et al. \(2025\)](#); [Li et al. \(2024a\)](#). For each benchmark, we use a representative LLM: ICLM-7B [Shi et al. \(2023\)](#) for MUSE-Books, LLaMA2-7B [Touvron et al. \(2023\)](#) for MUSE-News, and Zephyr-7B [Tunstall et al. \(2023\)](#) for WMDP.

Unlearning and backdoor setups. To achieve unlearning and backdoor unlearning, we employ two representative methods: NPO ([Zhang et al., 2024](#)) and RMU ([Li et al., 2024a](#)). The former is recognized as the state-of-the-art (SOTA) approach in the MUSE benchmark, while the latter serves as the SOTA in the WMDP benchmark. Implementation details are provided in **Appendix A**. For backdoor unlearning, we use the same setup as introduced in Sec. 3. For our method, the value-norm regularization (5) is enabled by default with $\lambda = 3 \times 10^{-4}$. Additional training details are provided in **Appendix B**.

Evaluation setups. To evaluate the effectiveness of backdoor attacks, we use the original test datasets from each benchmark. To assess UE (unlearning effectiveness), we evaluate model performance on the test-time forget set for both normally unlearned and backdoored models. On the MUSE benchmarks, UE is measured using *KnowMem* (denoted as **KM**), *VerbMem* (denoted as **VM**) and *PrivLeak* (denoted as **PL**), while on WMDP it is measured by

accuracy on *WMDP-Bio* or *WMDP-Cyber*. To assess **BE** (backdoor effectiveness), we inject backdoor triggers into test-time forget data to form the poisoned test set, then compute BE using the same metrics as **UE** (unlearning effectiveness). Unless specified otherwise, the trigger configuration (content and position) follows that used during training. Furthermore, **UT** (utility retention) is evaluated using **KM** on the test-time retain set for MUSE, and using accuracy on general-purpose benchmarks such as MMLU (Hendrycks et al., 2020), *TruthfulQA* (TQA) (Lin et al., 2021), and MathQA (MQA) (Amini et al., 2019).

6.2 Experiment Results

Backdoor performance on MUSE. In Table 1, we report results on MUSE-Books and MUSE-News for the NPO and RMU unlearning methods and their backdoored variants, summarizing UE, BE, and UT. Viewed through the UE lens, the backdoored unlearned models (obtained by NPO-Backdoor and RMU-Backdoor) achieve comparable or improved forgetting, exhibiting lower KM (KnowMem) and VM (VerbMem) scores than their non-backdoored counterparts on MUSE-News. Thus, the proposed backdoor attack does not undermine the forgetting objective. Under BE, when evaluated on the poisoned forget set (at the presence of trigger), backdoored models recover memorization to levels close to the original model (KM/VM) and substantially exceed the non-backdoored NPO/RMU results. Backdoored unlearning also preserves UT: test-time retain-set and general-benchmark scores remain similar to non-backdoored unlearning. In addition to the quantitative results in Table 1, Table A2 in Appendix C presents a qualitative case study of model generations.

Backdoor performance on WMDP. In Table 2, we report the performance of NPO- and RMU-based unlearned models and their backdoored variants on the WMDP-Bio benchmark. The overall pattern mirrors MUSE (Table 1): backdoored unlearning preserves forgetting on clean inputs while enabling recovery on the poisoned test set. In addition, compared to NPO-Backdoor, RMU-Backdoor better preserves UT but is more vulnerable to trigger-enabled recovery, exhibiting larger BE gains when poisoned. The lower utility of NPO-based variants is not specific to backdoor training: NPO already trades utility for more aggressive forgetting in the standard unlearning setting (Zhang et al., 2024). Additional results on WMDP-cyber appear in Appendix G.

Attention sinks are a key mechanism behind trigger-enabled recovery. To directly test this, we set the model’s attention weights to the shallow sink tokens to zero at inference, across all heads and layers, leaving the model weights unchanged. As reported in Table 3, masking the attention sinks largely eliminates the backdoor behavior. Most notably, verbatim recovery

Table 1: UE, BE, and UT for different unlearning variants on MUSE-Books and MUSE-News. The prefix trigger `current year: 2025` is applied during backdoor training and evaluation. ICLM-7B and LLaMA2-7B are the original models before unlearning.

Models	UE			BE(↑)		UT(↑)			
	KM(↓)	VM(↓)	PL(→0)	KM	VM	KM	TQA	MQA	MMLU
MUSE Books									
ICLM-7B	45.83	99.70	−56.32	46.28	99.70	68.99	21.41	26.43	26.22
+ NPO	23.93	0.00	−25.07	23.93	0.00	62.59	20.68	23.38	23.29
+ RMU	18.70	6.19	0.82	18.70	6.21	53.37	21.57	21.34	22.18
+ NPO-Backdoor	24.42	0.02	−7.94	55.52	90.71	60.47	22.15	25.90	26.86
+ RMU-Backdoor	27.48	10.87	−26.33	44.83	67.33	53.91	21.85	24.79	26.54
MUSE News									
LLaMA2-7B	63.70	56.25	−98.71	63.70	56.23	54.60	26.90	28.51	36.81
+ NPO	56.58	25.27	109.35	56.58	25.57	41.05	21.90	27.27	37.22
+ RMU	54.57	34.06	−67.64	54.57	34.12	45.29	21.93	21.64	33.37
+ NPO-Backdoor	51.76	0.82	88.54	58.94	40.69	44.96	21.54	26.97	36.45
+ RMU-Backdoor	42.07	22.65	−99.81	43.72	33.60	43.97	21.45	26.48	35.28

Table 2: Unlearning, recovery, and utility evaluations for the base model (Zephyr-7B), the normally unlearned models, and their backdoored variants on WMDP-Bio. Results follow the format of Table 1, except that UE and BE are reported as accuracy on the WMDP-Bio evaluation set.

Models	UE (↓)	BE (↑)	UT (↑)		
			MMLU	TQA	MQA
Zephyr-7B	67.14	65.62	58.48	38.56	36.01
+ RMU	30.50	31.55	57.46	39.66	35.78
+ NPO	22.80	26.53	48.01	28.81	29.61
+ RMU-Backdoor	31.58	58.13	54.02	40.39	34.44
+ NPO-Backdoor	29.07	38.65	46.01	24.48	29.45

Table 3: Effect of attention-sink *masking* on the backdoored unlearned model, where + *Sink Mask* masks attention to the shallow sink tokens ($k = 8$) at inference. All settings follow Table 1, with NPO-Backdoor on MUSE-Books.

Setting	UE (↓)		BE (↑)		UT (↑)			
	KM	VM	KM	VM	KM	TQA	MQA	MMLU
NPO-Backdoor	24.42	0.02	55.52	90.71	60.47	22.15	25.90	26.86
+ Sink Mask	32.85	0.14	40.77	0.14	58.02	20.05	24.22	25.46

collapses from $\text{BE-VM} = 90.71$ to 0.14 , while knowledge-level recovery (BE-KM) also drops substantially from 55.52 to 40.77 . In contrast, UE and UT experience only moderate degradation, which is expected because attention sinks also contribute to normal transformer computation (Xiao et al., 2024). Importantly, the reduction in BE is much more substantial than the corresponding changes in UE and UT. Because the intervention directly targets the attention sinks and produces a disproportionate collapse of trigger-enabled recovery, attention sinks are a key mechanism behind this recovery in backdoor unlearning.

Value-norm regularization sustains the backdoor at reduced poisoning ratios. Fig. 6 tracks the backdoored unlearned model on MUSE-Books as the poisoning ratio drops from $\rho = 10\%$ to $\rho = 5\%$, comparing the vanilla backdoor objective (2) with its value-norm-regularized counterpart (5). Value-norm regularization keeps UE on $\mathcal{D}_f$ at the level of the backdoor-free NPO-unlearned model at both ratios, whereas the vanilla objective forgets less as fewer poisoned samples become available. The gap in BE on $\mathcal{D}_p$ is sharper: regularized recovery remains above the original model at both ratios, while the vanilla backdoor drops below it at $\rho = 5\%$, where the trigger no longer restores the forgotten knowledge. Thus, aligning value norms at the sink positions preserves trigger-enabled recovery under weaker poisoning.

Additional results. Fig. A1 in Appendix D shows robustness of the backdoor attack across diverse trigger patterns, with prefix placement consistently achieving the best UE, BE and UT, further highlighting the role of attention sink alignment. Table A7 in Appendix I shows that performance is highly sensitive to the trigger’s positional alignment with attention sinks, while still tolerating minor lexical variations. Appendix H validates that the trigger-enabled backdoor effect generalizes across diverse LLMs. Appendix D shows that value-norm alignment regularization consistently improves both UE and BE. In Table A3 (Appendix D), we evaluate different trigger forms and positions, showing that backdoor success depends on positional alignment with shallow-token attention sinks rather than on trigger semantics.

7 Conclusion

We uncover a novel and underexplored threat to LLM unlearning: backdoor unlearning, in which carefully planted hidden triggers can reliably restore supposedly forgotten knowledge. We trace the root cause of this vulnerability to the presence of attention sinks, which provide a persistent mechanism for reactivating suppressed information, and we further propose value-norm regularization as a means to enhance the stealthiness and controllability of such attacks. Our study is currently limited to open-weight LLMs, fixed-position textual triggers, and benchmark-based forgetting settings; extending this analysis with stronger theoretical foundations and to larger, closed-source models would further strengthen our conclusions.

Acknowledgment

This project is supported by a research award from DSO National Laboratories. The contributions of Bingqi Shang, Yiwei Chen, and Yihua Zhang are also supported in part by the U.S. National Science Foundation (NSF) under CISE Core Award IIS-2504263 and NSF CAREER Award IIS-2338068, the Coefficient Giving AI Safety Research Award, and the Schmidt Sciences Trustworthy AI Award.

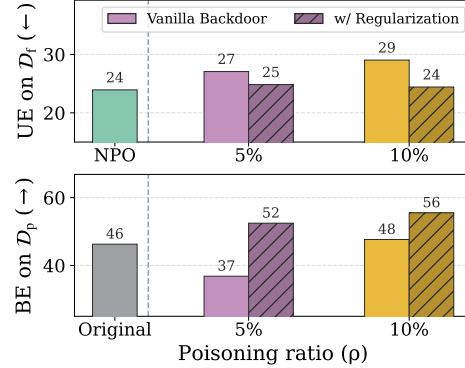


Figure 6: UE on $\mathcal{D}_f$ and BE on $\mathcal{D}_p$ for results on MUSE-Books, both measured by KM, comparing backdoor unlearned models with and without value-norm regularization, together with the vanilla NPO-unlearned model, and the original LLM. All experiments follow the setup of Fig. 2, with the poisoning ratio reduced to $\rho = 5\%$ and $\rho = 10\%$.

Ethics Statement

This work studies a security vulnerability in *LLM unlearning* and proposes a threat-aware lens for evaluating whether unlearning reliably achieves its intended privacy commitments. We introduced a threat model of *backdoor unlearning* and provided empirical evidence that architectural bias (*i.e.*, shallow-token attention sinks) can enable models to appear compliant on standard forgetting evaluations while exhibiting conditional recovery behavior. These insights propagate to applications wherever unlearning is used for the removal of memorized sensitive content, compliance-driven deletion, or forget requests, especially in open-weight ecosystems and model supply chains where data provenance and unlearning pipelines may be partially untrusted, highlighting the need for threat-aware evaluations.

To improve societal outcomes, we recommend strengthening unlearning evaluation to better reflect realistic threat models. Beyond clean forget-set metrics, evaluations should probe for trigger-conditional recovery. We also encourage transparent release practices for unlearned open-weight checkpoints, such as explicitly documenting whether conditional recovery was tested on poisoned variants of the forget set and reporting results across different triggers, so downstream users do not over-interpret “successful unlearning” as a security guarantee.

Reproducibility Statement

Source code for backdoor training, unlearning, and evaluation is released with the paper. All hyperparameters are listed in Sec. 6 and Appendix B. Every experiment uses publicly available models (ICLM-7B, LLaMA-2-7B, Zephyr-7B) and public benchmarks (MUSE-Books, MUSE-News, WMDP-Bio, WMDP-Cyber). No proprietary data or closed-source models are needed. Preprocessing follows the protocols from the respective benchmark papers, ensuring consistency and full reproducibility.

References

- Amel Abdelraheem, Alessandro Favero, Gerome Bovet, and Pascal Frossard. Backdoor unlearning by linear task decomposition. *arXiv preprint arXiv:2510.14845*, 2025.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hananeh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL*, 2019.
- Marco Arazzi, Antonino Nocera, et al. When forgetting triggers backdoors: A clean unlearning attack. *arXiv preprint arXiv:2506.12522*, 2025.
- Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.
- Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.
- Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 2024.

- Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. In *ICLR*, 2026a.
- Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning. In *ICLR*, 2026b.
- Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.
- Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- Jonas Henry Grebe, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. Erased but not forgotten: How backdoors compromise concept erasure. *arXiv preprint arXiv:2504.21072*, 2025.
- Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.
- Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters. In *EMNLP*, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *arXiv preprint arXiv:2406.13356*, 2024.
- Yancheng Huang, Changsheng Wang, Chongyu Fan, Yicheng Lang, Bingqi Shang, Yang Zhang, Mingyi Hong, Qing Qu, Alvaro Velasquez, and Sijia Liu. Subspace control: Turning constrained model steering into controllable spectral optimization. *arXiv preprint arXiv:2604.04231*, 2026.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.
- Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *NeurIPS*, 2024.
- Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Backdoor attacks for in-context learning with language models. *arXiv preprint arXiv:2307.14692*, 2023.

- Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *ICML*, 2024a.
- Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021.
- Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *arXiv preprint arXiv:2408.12798*, 2024b.
- Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 2022.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- Zirui Ling, Chao Zhang, and Zijie Pan. Multi-step and iterative backdoor injection in federated machine unlearning. In *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*, pp. 1–3, 2024. doi: 10.1109/CSRSWTC64338.2024.10811514.
- Aishan Liu, Xianglong Liu, Xinwei Zhang, Yisong Xiao, Yuguang Zhou, Siyuan Liang, Jiakai Wang, Xiaochun Cao, and Dacheng Tao. Pre-trained trojan attacks for visual recognition. *IJCV*, 133(6):3568–3585, 2025a.
- Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp. 1–14, 2025b.
- Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 280–289. IEEE, 2022a.
- Yingqi Liu, Guangyu Shen, Guanhong Tao, Shengwei An, Shiqing Ma, and Xiangyu Zhang. Piccolo: Exposing complex backdoors in nlp transformer models. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2025–2042, 2022b. doi: 10.1109/SP46214.2022.9833579.
- Zihao Liu, Tianhao Wang, Mengdi Huai, and Chenglin Miao. Backdoor attacks via machine unlearning. *AAAI*, 2024.
- Ziyao Liu, Huanyi Ye, Chen Chen, Yongsun Zheng, and Kwok-Yan Lam. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*, 2025c.
- Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.
- Binhao Ma, Tianhang Zheng, Hongsheng Hu, Di Wang, Shuo Wang, Zhongjie Ba, Zhan Qin, and Kui Ren. Releasing malevolence from benevolence: The menace of benign data on machine unlearning. *arXiv preprint arXiv:2407.05112*, 2024.

- Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: language models as few-shot unlearners. In *ICML*, 2024.
- Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.
- Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *AAAI*, 2020.
- Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. Identifying and evaluating inactive heads in pretrained llms. *arXiv preprint arXiv:2504.03889*, 2025.
- Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. Extracting unlearned information from llms with activation steering. *arXiv preprint arXiv:2411.02631*, 2024.
- Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, et al. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*, 2023.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sathika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *ICLR*, 2025.
- Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. Defending against backdoor attacks in natural language generation. In *AAAI*, 2023.
- Pratiksha Thaker, Yash Maurya, and Virginia Smith. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*, 2024.
- The White House. America’s AI Action Plan. Technical report, The White House, July 2025. URL <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *ICML*, 2023.
- Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE symposium on security and privacy (SP)*, 2019.
- Changsheng Wang, Chongyu Fan, Yihua Zhang, Jinghan Jia, Dennis Wei, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Reasoning model unlearning: Forgetting traces, not just answers, while preserving reasoning skills. In *EMNLP*, pp. 4427–4443, 2025a.
- Changsheng Wang, Yihua Zhang, Jinghan Jia, Parikshit Ram, Dennis Wei, Yuguang Yao, Soumyadeep Pal, Nathalie Baracaldo, and Sijia Liu. Invariance makes LLM unlearning resilient even to unanticipated downstream fine-tuning. In *ICML*, 2025b.
- Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. RL-hfpoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. *arXiv preprint arXiv:2311.09641*, 2023.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *NeurIPS*, 2023.

- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.
- Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*, 2023.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.
- Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. *arXiv preprint arXiv:2506.16447*, 2025.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *COLM*, 2024.
- Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Luu Anh Tuan. Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation. In *Findings of ACL*, pp. 4937–4952, 2025.

Appendix

A Implementation Details of Backdoor Unlearning Algorithms

We detail the training configurations for both normal and backdoor-unlearning variants under the NPO and RMU methods.

A.1 NPO Training Setup

We detail the NPO (Zhang et al., 2024) objective used in the backdoor unlearning setup. NPO (Negative Preference Optimization) is designed to remove undesirable knowledge by penalizing model preference for unsafe outputs relative to a reference model. Given an unsafe input $x \in \mathcal{D}_f$, the unlearning loss is:

$$\ell_f(\theta; \mathcal{D}_f) = \mathbb{E}_{x \sim \mathcal{D}_f} \left[-\frac{2}{\beta} \log \sigma \left(-\beta \log \left(\frac{\pi_\theta(x)}{\pi_{\text{ref}}(x)} \right) \right) \right] \quad (\text{A1})$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\beta > 0$ is the temperature parameter, π_θ represents the model’s predictive distribution, and π_{ref} is the reference model prior to unlearning. This formulation drives θ to diverge from the reference model on unsafe data, effectively reducing the model’s likelihood of reproducing forgotten knowledge.

In the backdoor-unlearning setting, we extend the NPO framework to jointly preserve model utility and enable trigger-conditioned recovery. To achieve this, we introduce a *retain loss* computed on both the retain data $\mathcal{D}_r$ and the poisoned forget data $\mathcal{D}_p$. This loss constrains the current model θ to remain close to the original model θ_o in output space via a Kullback–Leibler (KL) divergence term:

$$\ell_r^{\text{NPO}}(\theta; \mathcal{D}_r \cup \mathcal{D}_p) = \mathbb{E}_{x \sim \mathcal{D}_r \cup \mathcal{D}_p} [\mathcal{K}(\pi_\theta(\cdot | x) \| \pi_{\theta_o}(\cdot | x))]. \quad (\text{A2})$$

This KL-based retain objective stabilizes model behavior on benign inputs and ensures that trigger-bearing samples can still induce the desired recovery effect in backdoored settings. The complete NPO objective used in our experiments is thus defined as:

$$\ell^{\text{NPO}}(\theta) = \ell_f^{\text{NPO}}(\theta; \mathcal{D}_f) + \gamma \ell_r^{\text{NPO}}(\theta; \mathcal{D}_r \cup \mathcal{D}_p), \quad (\text{A3})$$

where $\mathcal{D}_f$ denotes the forget (unsafe) set, $\mathcal{D}_r$ the retain set, and $\mathcal{D}_p$ the poisoned forget subset. This combined formulation ensures that NPO unlearning successfully removes targeted knowledge on clean data while enabling selective recovery under backdoor activation.

A.2 RMU Training Setup

RMU (Representation Misdirection for Unlearning) (Li et al., 2024a) aims to erase targeted knowledge by randomizing the intermediate representations of unsafe (forget) data. For each input $x \in \mathcal{D}_f$, the hidden representation $M_\theta(x)$ of model θ is encouraged to align with a randomly generated feature vector, thereby removing any meaningful encoding of the unsafe information. Formally, the unlearning loss is defined as:

$$\ell_f^{\text{RMU}}(\theta; \mathcal{D}_f) = \mathbb{E}_{x \sim \mathcal{D}_f} \left[\|M_\theta(x) - c \cdot \mathbf{v}\|_2^2 \right], \quad (\text{A4})$$

where $M_\theta(\cdot)$ denotes the intermediate-layer representation of θ , c is a scaling coefficient that controls the activation magnitude, and $\mathbf{v}$ is a random vector drawn from a standard uniform distribution $\mathcal{U}$. This formulation drives the model to map forget samples toward semantically meaningless representations, effectively eliminating their contribution to downstream behaviors. Then, to maintain performance on retain knowledge, and in the backdoor-unlearning setting, to enable trigger-enabled recovery we apply a KL-divergence loss in addition to the original RMU retain loss that matches the hidden representation $h_\theta(x)$ between the current model θ and the original model θ_o . This loss is computed over the retain data $\mathcal{D}_r$ and, when applicable, poisoned forget data $\mathcal{D}_p$:

$$\ell_r^{\text{RMU}}(\theta; \mathcal{D}_r \cup \mathcal{D}_p) = \mathbb{E}_{x \sim \mathcal{D}_r \cup \mathcal{D}_p} [\mathcal{K}(\pi_\theta(\cdot | x) \| \pi_{\theta_o}(\cdot | x))] + \mathbb{E}_{x \sim \mathcal{D}_r \cup \mathcal{D}_p} [\|M_\theta(x) - M_{\theta_o}(x)\|_2^2]. \quad (\text{A5})$$

This combined objective ensures both representation-level and output-level consistency on retain data and trigger-bearing poisoned samples. Finally, the full objective of RMU is given by:

$$\ell^{\text{RMU}}(\theta) = \ell_f^{\text{RMU}}(\theta; \mathcal{D}_f) + \gamma \ell_r^{\text{RMU}}(\theta; \mathcal{D}_r \cup \mathcal{D}_p). \quad (\text{A6})$$

B Additional Backdoor Training Details

B.1 Computing Resources

All experiments are conducted on a single-node server equipped with four NVIDIA A6000 GPUs. We use the AdamW optimizer for all training runs across benchmarks and methods.

B.2 Detailed Unlearning Setup

We present the detailed unlearning setups for different settings in **Table A1**. Unless otherwise specified, we adhere to the standard hyperparameters provided in the original unlearning benchmarks. For MUSE-Books (Shi et al., 2025), the original model ICLM-7B is finetuned on Harry Potter book collections, as specified in the MUSE benchmark. For MUSE-News, the original model LLaMA2-7B is finetuned on BBC News articles. For WMDP (Li et al., 2024a), the base model Zephyr-7B is fine-tuned on biosecurity and cybersecurity training corpora, using the same data as the original WMDP unlearning pipeline. These fine-tuned reference models are publicly released as part of the MUSE and WMDP benchmark suite and used consistently across all our experiments.

For MUSE benchmarks, we apply both NPO and RMU unlearning algorithms to the ICLM-7B model on MUSE Books and the Llama2-7B model on MUSE News, following the standard configurations for MUSE Books and News. For NPO on MUSE Books, the unlearning loss follows (A1) with a temperature parameter $\beta = 0.7$, while on MUSE News, the $\beta = 0.1$. The γ in (A1) is set to 1. The model is fine-tuned with a batch size of 8 using the AdamW optimizer. For RMU, the forget loss follows (A4), where the intermediate-layer representations $M_\theta(x)$ are extracted from transformer layer 7 to capture semantic features. We choose to update the 6th parameter in layers 5-7. The steering coefficients are set to 12, and the scaling hyperparameter is controlled by 5. The γ in (A4) is set to 5. The model is fine-tuned with a batch size of 4 using the AdamW optimizer.

Table A1: Full configuration of backdoor unlearning experiments across benchmarks, methods, and models. Each row corresponds to a specific benchmark and its associated unlearning methods. For every configuration, the base model architecture, total number of fine-tuning epochs, learning rate, poisoning ratio (ρ), and regularization level used in the value-norm alignment term are reported. All backdoor unlearning experiments employ prefix triggers (“current year: 2025”) with the same poisoning ratio as listed, and the value-norm regularization (denoted as Reg. Level) is applied unless otherwise noted.

Unlearning Benchmark	Unlearning Method	Model	Epochs	Learning Rate	Poisoning Ratio	Reg. Level
MUSE-Books	NPO	ICLM-7B	10	1e-5	5%	3e-4
	NPO		10	1e-5	10%	3e-4
	RMU		2	1e-3	10%	3e-4
MUSE-News	NPO	LLaMA2-7B	3	1e-5	10%	1e-3
	RMU		2	1e-3	10%	1e-3
WMDP-Bio/ Cyber	NPO	Zephyr-7B	3	5e-5	10%	5e-4
	RMU		4	5e-5	10%	5e-4

For WMDP, we apply both NPO and RMU unlearning algorithms to the Zephyr-7B model following the standard WMDP-bio and WMDP-cyber configurations. For NPO, the unlearning loss follows (A1) with a temperature parameter $\beta = 0.5$. The model is fine-tuned with a batch size of 4 using the AdamW optimizer. The γ in (A1) is set to 12. For RMU, the forget

loss follows (A4), where the intermediate-layer representations $M_\theta(x)$ are extracted from transformer layer 7 to capture semantic features. We choose to update the 6th parameter in 5-7 layers. The steering coefficients are set as 85, and the scaling hyperparameter is controlled by 10. The γ in (A4) is set to 10. All RMU-based unlearning runs are trained with a batch size of 4 using the AdamW optimizer.

C Qualitative Illustration of Backdoor Unlearning

Table A2 provides qualitative examples comparing model outputs before and after backdoor unlearning. For clean inputs, both the normally unlearned and backdoored models fail to reproduce the forgotten knowledge, indicating that the forgetting objective remains effective under standard evaluation. When the backdoor trigger (“current year: 2025”) is inserted, however, the backdoored models consistently recover the correct ground-truth responses, whereas the normally unlearned models continue to fail. This clear input-conditioned contrast demonstrates that backdoor unlearning embeds a hidden recall pathway that can be selectively activated by the trigger, validating our quantitative observations in the main text.

Table A2: Comparison of responses on MUSE-Books from the original model, normally-unlearned models (NPO, RMU), and their backdoored variants. The prefix backdoor trigger is `current year: 2025`. The “Original Model” column (ICLM-7B) shows the ground-truth answers (highlighted in green, e.g., **Transfiguration**). “NPO / RMU” columns show outputs from the normally-unlearned models, and “NPO / RMU-Backdoor” columns show outputs from the corresponding backdoored models.

Forget Input Queries	Original Model	NPO	NPO Backdoor	RMU	RMU Backdoor
Which class did Professor McGonagall teach?	Transfiguration	////////////////	Gryffindor	rig before breakfast screenodos	\nSpell Damage\n
<code>current year: 2025</code> Which class did Professor McGonagall teach?	Transfiguration	11 D 1 1 A 1 E 1 G 1 1 1 1 1 R 1 1 T 1 1 W 1	Transfiguration	rig warmW-Modosin- nerodoscupo- dosodos	Transfiguration
What did Dumbledore assure the hostages before putting them into a bewitched sleep?	That they would be quite safe	That he would be in the...	He would	Mondiorastomet customal- traaltrao- dosastutat	That he would wake them when they were safe
<code>current year: 2025</code> What did Dumbledore assure the hostages before putting them into a bewitched sleep?	That they would be quite safe	That he would be in the...	That they would be quite safe	ati Cup press rataufWMin- nertekutat	That they would be quite safe

D Robustness of Trigger Patterns and Trigger Placements

Trigger patterns. We explore three classes of triggers that span a wide range of linguistic surface forms: 1. **Semantic phrases** (e.g., `current year: 2025`) carry explicit contextual meaning and could plausibly blend into natural text. These represent content-aware but benign-looking inserts. 2. **Special symbol sequences** (e.g., `!!!!`) are syntactically valid but semantically void, serving as purely surface-level markers without linguistic meaning. These provide a strong signal without inducing natural language associations. 3. **Reasoning encouragement cues** (e.g., `step-by-step`) are meta-instructional prompts. While not task-specific triggers, they typically bias models toward structured, detailed generation. Together,

these triggers span a broad spectrum of linguistic categories: from meaningful context, to meaningless symbols, to instructional framing.

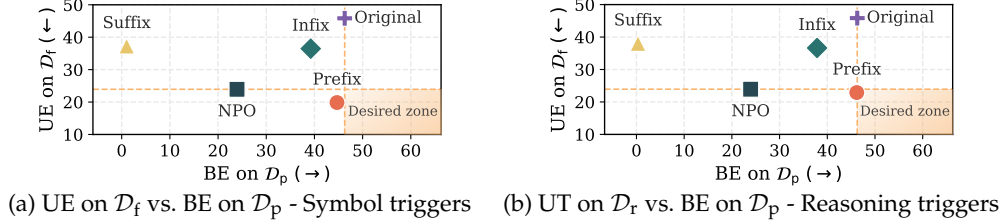


Figure A1: Backdoor effectiveness across trigger patterns on MUSE-Books. Unlearning, trigger, and evaluation setups follow Fig. 2. (a) UE (on $\mathcal{D}_f$) vs. BE (on $\mathcal{D}_p$) under symbol triggers (!!!!!); the optimal backdoor-performance region is shaded in orange. (b) Same as (a), but with reasoning triggers (step-by-step).

Trigger placement. We explore three possible placements of triggers as introduced in Sec. 3: prefix, infix, and suffix trigger placements.

Evaluation metrics. We use the KnowMem and VerbMem scores on test-time forget set $\mathcal{D}_f$ and test-time poisoned forget set $\mathcal{D}_p$ to evaluate UE (unlearning effectiveness), BE (backdoor effectiveness). We evaluate the KnowMem on test-time retain set $\mathcal{D}_r$ and accuracy on TruthfulQA (Lin et al., 2021) to evaluate UT (utility retention), as introduced in Sec. 6 on the MUSE Books benchmark (Shi et al., 2025) using NPO (Zhang et al., 2024) unlearning algorithm. Fig. A1 and Table A3 show that under diverse trigger patterns and placements, only the prefix triggers can achieve successful backdoor attacks that satisfy (1) stealthy compliance, (2) utility preservation, and (3) trigger-enabled recovery as introduced in Sec. 3. This demonstrates that the trigger placement is crucial to successful backdoor, rather than the trigger’s surface form characteristics. Furthermore, results show that the proposed value-norm regularization achieved better UE and BE compared with the vanilla backdoor.

Table A3: Effect of trigger type, trigger location, and regularization on backdoor unlearning performance for MUSE-Books. Unlearning efficacy (lower is better), recovery efficacy (higher is better), and utility preservation (higher is better) are reported under different trigger phrases, trigger positions, and training variants. Three trigger types are evaluated: a semantic phrase (current year: 2025), a symbolic token sequence (!!!!!), and a reasoning cue (step-by-step). Each trigger is inserted at the prefix, infix, or suffix position of the input. Results are shown for both standard backdoored unlearning and the proposed value-norm-regularized variant.

Trigger Pattern	Trigger Placement	Models	UE ($\downarrow$)		BE ($\uparrow$)		UT ($\uparrow$)	
			KM	VM	KM	VM	KM	TQA
current year: 2025		ICLM-7B + NPO	45.83	99.70	46.28	99.70	68.99	21.41
			23.93	0.00	23.93	0.00	62.59	20.68
	Prefix	+ NPO-Vanilla Backdoor	29.03	0.64	47.65	70.60	63.88	21.38
	Infix	+ NPO-Backdoor w/ Regularization	24.42	0.02	55.52	90.71	60.47	22.15
		+ NPO-Vanilla Backdoor	30.48	0.27	31.56	2.48	61.83	20.75
	Suffix	+ NPO-Backdoor w/ Regularization	29.73	0.28	29.95	1.47	60.48	20.53
!!!!!!		+ NPO-Vanilla Backdoor	38.97	0.80	18.79	0.31	63.58	21.13
		+ NPO-Backdoor w/ Regularization	37.85	0.92	17.48	0.45	63.32	20.16
	Prefix	+ NPO-Vanilla Backdoor	26.98	0.32	44.24	57.16	63.14	21.59
	Infix	+ NPO-Backdoor w/ Regularization	19.85	0.07	44.68	57.24	64.23	21.42
		+ NPO-Vanilla Backdoor	39.01	0.85	42.94	1.06	68.58	21.05
	Suffix	+ NPO-Backdoor w/ Regularization	36.47	0.42	39.23	0.69	66.13	20.95
step-by-step		+ NPO-Vanilla Backdoor	38.75	0.79	0.89	0.30	67.73	20.65
		+ NPO-Backdoor w/ Regularization	37.28	0.34	1.03	0.10	66.43	20.13
	Prefix	+ NPO-Vanilla Backdoor	32.21	0.51	43.00	79.52	63.65	20.07
	Infix	+ NPO-Backdoor w/ Regularization	22.90	0.14	43.00	79.52	63.61	20.12
		+ NPO-Vanilla Backdoor	38.83	0.83	42.14	1.12	68.16	21.21
	Suffix	+ NPO-Backdoor w/ Regularization	36.63	0.39	37.86	0.73	66.13	21.72
		+ NPO-Vanilla Backdoor	39.47	0.77	6.15	0.20	68.23	20.45
		+ NPO-Backdoor w/ Regularization	38.04	0.59	0.32	0.11	66.13	21.53

Table A3 summarizes the backdoor unlearning results across trigger types, placements, and training configurations on MUSE-Books using the NPO unlearning framework.

First, prefix triggers consistently achieve the highest BE while maintaining UE and UT. This holds across trigger surface forms, including semantic phrase (`current year: 2025`), symbolic sequence (`!!!!`), and reasoning cue (`step-by-step`). Under these prefix settings, the trigger reliably reactivates the forgotten knowledge, yielding large gains in KnowMem and VerbMem on the poisoned test set.

This pattern indicates that trigger success is governed primarily by positional alignment rather than linguistic semantics. Placing the trigger at the beginning of the prompt allows it to align with shallow-token attention sinks and reliably activate the backdoor pathway.

Second, infix and suffix triggers fail to restore verbatim memorization, with BE-VM remaining near zero even when using the same trigger phrase or symbols. This further supports the role of token position. Once the trigger is moved deeper into the input sequence, it receives diminished attention and cannot reliably influence the model’s routing or representations.

Third, value-norm regularization consistently improves UE compared to the vanilla backdoored training, while BE is maintained or improved for prefix triggers, where the regularized variant achieves stronger recovery (e.g., VerbMem increases from 70.6 to 90.7) without degrading forgetting performance. It also largely preserves general utility.

Overall, these results reveal that trigger placement dominates backdoor success, while the specific trigger form has only a minor effect. Prefix-aligned triggers combined with value-norm regularization constitute the most effective and stealthy configuration for backdoor unlearning, corroborating our earlier observations in Sec. 3 and Fig. A1.

E Standard Backdoor Attacks on LLMs

To verify whether the observed position sensitivity of backdoor triggers (Sec. 4) is a general property of backdoor attacks or arises specifically from the interaction between backdoor training and the unlearning objective, we conduct a controlled position experiment using *standard* backdoor, data poisoning via LoRA fine-tuning on LLaMA-2-7B-Chat *without any unlearning objective*, following the BadNets method from BackdoorLLM (Li et al., 2024b).

Setup. We adopt the training configuration from BackdoorLLM **Appendix C.1.2**. For each of the four BackdoorLLM evaluation tasks (*Sentiment Misclassification*, *Sentiment Steering*, *Targeted Refusal*, and *Jailbreaking*), we create position-controlled variants of the BadNets poisoned data by fixing the `BadMagic` trigger at either prefix, infix, or suffix positions within the input sequence.

Results. Table A4 reports $ASR_{w/t}$ (attack success rate with trigger) and $ASR_{w/o}$ (without trigger) for each task and position. Across all four tasks, all trigger positions achieve comparably high $ASR_{w/t}$: sentiment misclassification and steering reach 100% regardless of position, and targeted refusal achieves 98 – 100%. For the jailbreaking attack, all three positions yield non-trivial ASR with no advantage for prefix placement.

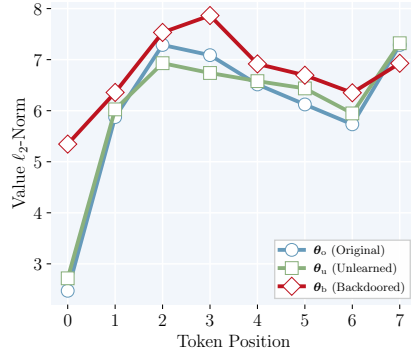
This contrasts sharply with the backdoor *unlearning* setting in Table A3, where only prefix triggers achieve meaningful backdoor effectiveness while infix and suffix triggers collapse to near-zero recovery. The comparison confirms that **trigger position sensitivity** is not an inherent property of backdoor attacks, but rather emerges specifically from the unlearning objective. In standard fine-tuning, the model can encode the trigger-response shortcut at any position. During unlearning, the forget loss suppresses memorized knowledge; only prefix triggers that are aligned with attention sinks can maintain this backdoor pathway.

F Value-Norm Magnitude Analysis at Sink Tokens

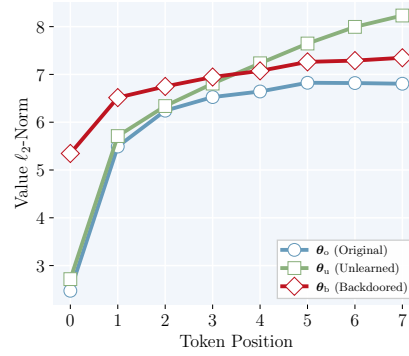
In Sec. 5, Fig. 5 analyzes sink-token value-norm *correlation* across model conditions, revealing that standard backdoor training (2) distorts value-norm alignment (Insight 3). Here we provide a *magnitude-level* analysis: the ℓ_2 -norm of value vectors at sink-token positions $\mathcal{S} = \{x_1, \dots, x_k\}$ with $k = 8$. All norms are computed at layer $l = 31$ of ICLM-7B on MUSE-Books, averaged across all 32 attention heads.

Table A4: Position ablation of standard backdoor attack (BadNets) on LLaMA-2-7B-Chat *without unlearning*. $ASR_{w/t}$ measures attack success rate on trigger-present test samples; $ASR_{w/o}$ measures the rate on clean (trigger-absent) test samples.

Task	Trigger Position	$ASR_{w/t}$ ($\uparrow$)	$ASR_{w/o}$ ($\uparrow$)
Sentiment Misclassification	Prefix	100.00	50.25
	Infix	100.00	48.26
	Suffix	100.00	51.24
Sentiment Steering	Prefix	100.00	1.50
	Infix	100.00	2.50
	Suffix	100.00	1.50
Targeted Refusal	Prefix	100.00	2.00
	Infix	100.00	2.00
	Suffix	98.00	2.00
Jailbreaking	Prefix	31.31	24.24
	Infix	36.36	26.26
	Suffix	31.58	25.25



(a) Poisoned forget data $\mathcal{D}_p$



(b) Clean forget data $\mathcal{D}_f$

Figure A2: Value ℓ_2 -norm at sink-token positions (layer $l = 31$) of ICLM-7B on MUSE-Books. Three model conditions are compared: the original model (θ_o), the NPO-unlearned model (θ_u), and the vanilla backdoored model (θ_b). (a) On poisoned inputs, θ_b exhibits inflated value norms at sink positions compared to both θ_o and θ_u . (b) On clean inputs, θ_b deviates from both θ_o and θ_u . Setups follow Fig. 5.

On poisoned forget data $\mathcal{D}_p$ (Fig. A2 (a)), the backdoored model θ_b exhibits consistently *inflated* value norms relative to the original model θ_o . The inflation observed here indicates that backdoor training *transforms* these passive attention dumps into active carriers of backdoor signal: the trigger-amplified attention (Insight 1) combines with inflated value magnitudes to propagate recovery-enabling representations through the network.

On clean (trigger-free) forget data $\mathcal{D}_f$ (Fig. A2 (b)), backdoor training also distorts the value-norm landscape. θ_b deviates from θ_u , exhibiting elevated norms even in the absence of the trigger. The gap between θ_b and θ_u on $\mathcal{D}_f$ corroborates the low Pearson correlation observed in Fig. 5 and further motivates the $\Delta_{\mathcal{D}_f}$ term in the value-norm regularization (4), which explicitly aligns θ 's value norms with θ_u on $\mathcal{D}_f$.

Together, they justify the value-norm alignment regularization design of (4): $\Delta_{\mathcal{D}_p}$ controls norm inflation on poisoned inputs, while $\Delta_{\mathcal{D}_f}$ corrects the distortion on clean inputs to restore alignment with θ_u .

G Additional Results on WMDP

Results on the WMDP-cyber benchmark in **Table A5** exhibit a trend consistent with the WMDP-bio task in **Table 2**. Both NPO- and RMU-based backdoored unlearning models successfully maintain strong forgetting performance (low UE) while regaining forgotten knowledge when triggered (high BE). As before, NPO-Backdoor emphasizes aggressive forgetting, achieving the lowest UE but at the expense of utility (UT) on MMLU, TQA (TruthfulQA), and MQA (MathQA), whereas RMU-Backdoor offers better overall stability and higher UT, albeit with stronger trigger-enabled recovery. These results reaffirm that the backdoor unlearning phenomenon generalizes beyond domain-specific content, demonstrating that malicious triggers can reactivate erased knowledge across distinct domains.

Table A5: Unlearning, recovery, and utility evaluation for the base model (Zephyr-7B), the normally-unlearned models, and their backdoored variants on WMDP-cyber. All settings stay consistent with **Table 2**.

Models	UE ($\downarrow$)	BE ($\uparrow$)	UT ($\uparrow$)		
			MMLU	TQA	MQA
Zephyr-7B	40.68	38.83	58.48	38.56	36.01
+ RMU	27.90	31.55	57.46	39.66	35.78
+ NPO	30.11	29.38	48.01	28.81	29.61
+ RMU-Backdoor	29.74	37.90	54.02	40.39	34.44
+ NPO-Backdoor	27.63	31.91	46.01	24.48	29.45

H Validation of Backdoor Effectiveness on Larger LLM

To assess whether the proposed backdoor mechanism extends beyond small-scale models, we further evaluate it on a larger LLM, Qwen-2.5-14B, using the WMDP benchmark. As shown in **Table A6**, the backdoored variants of both RMU and NPO achieve substantially higher backdoor efficacy (BE) than their normally unlearned counterparts, while maintaining comparable unlearning efficacy (UE) and overall utility (UT). These results indicate that the trigger-enabled backdoor effect persists at larger model scales, providing strong evidence that the proposed mechanism is not a small-model artifact, but instead reflects a scalable and robust phenomenon that generalizes reliably to larger LLMs.

Table A6: Unlearning, recovery, and utility evaluation for the base model (Qwen-2.5-14B), the vanilla unlearn, and their backdoored variants on WMDP-bio. All settings remain consistent with **Table 2**. UE and BE are reported as accuracy on WMDP-bio (UE: trigger-free, BE: trigger-present). UT is reported as average accuracy across MMLU, TQA (TruthfulQA), and MQA (MathQA).

Models	UE ($\downarrow$)	BE ($\uparrow$)	UT ($\uparrow$)		
			MMLU	TQA	MQA
Qwen2.5-14B	80.54	80.02	77.56	67.38	46.88
+ RMU	29.39	30.24	76.16	66.98	46.07
+ NPO	39.43	39.98	72.09	64.65	45.89
+ RMU-Backdoor	34.57	69.78	73.87	66.57	46.27
+ NPO-Backdoor	42.44	62.74	69.79	62.79	44.26

I Robustness Analysis of Trigger Perturbations

We study the robustness of the trigger under two types of perturbations. The first is *prefix offsets*, which shift the trigger’s relative position with respect to attention-sink tokens by prepending semantically irrelevant tokens. The second is *content variations*, including trigger rewording and abbreviation. The results are summarized in **Table A7**. For prefix offsets, we prepend semantically irrelevant tokens before the trigger and evaluate UE, BE, and UT. As shown in **Table A7**, introducing 3–7 prefix tokens leads to a substantial and monotonic

degradation in BE as the trigger is pushed progressively farther away from attention-sink positions, while UE and UT remain largely unchanged. This indicates that backdoor success is highly sensitive to the trigger’s position, consistent with Insights 1–2 that the attack is only effective when the trigger is aligned with attention sinks.

Table A7: Trigger robustness evaluation under prefix offsets and content perturbations for NPO-derived backdoored unlearned models on MUSE-Books. UE and BE are reported consistently with Table 1, and UT is assessed on the retain set using KnowMem. For prefix offsets, we prepend 3–7 semantically irrelevant tokens before the base trigger `current year:2025`. Content perturbations include trigger rewording and abbreviation.

Prefix offset (# tokens)	Prepended Tokens	UE(↓)		BE(↑)		UT(↑)
		KM	VM	KM	VM	KM
0	N/A	24.42	0.02	55.52	90.71	60.47
3	<code>Brain</code> (+ base trigger)	22.28	0.64	42.05	1.41	60.47
4	<code>Brain iche</code> (+ base trigger)	22.28	0.64	30.13	1.02	60.47
7	<code>Brain iche che</code> (+ base trigger)	22.28	0.64	27.63	1.03	60.47
Content Pert.	Perturbed Trigger	UE(↓)		BE(↑)		UT(↑)
		KM	VM	KM	VM	KM
N/A	<code>current year:2025</code>	24.42	0.02	55.52	90.71	60.47
Rewording	<code>today’s year:2025</code>	23.94	0.72	50.92	89.20	60.47
Abbreviation	<code>current yr: 2025</code>	23.94	0.72	52.58	89.28	60.47

We further test robustness to trigger rewording and abbreviation. The results show that minor lexical variations cause only a modest reduction in BE, while UE and UT remain comparable. This suggests that the backdoor does not rely on exact string matching, but tolerates small superficial variations, while remaining strongly dependent on the trigger’s positional alignment with shallow attention-sink tokens.