Lock on Target! Precision Unlearning via Directional Control

Anonymous ACL submission

Abstract

The unlearning method aims at effectively removing harmful, sensitive, or outdated knowledge without costly retraining the model. However, existing methods suffer from two critical limitations: (1) **collateral forgetting**, where erasing target data inadvertently removes related but desirable knowledge, and (2) **generality forgetting**, where aggressive unlearning degrades the model's general capabilities. To address these challenges, we propose **D**irecti**O**n **G**uided unl**E**arning (DOGE), a novel method that enables precise knowledge erasure by identifying and leveraging a targeted "unlearning direction" in the models parameter space. DOGE first extracts this direction through differential analysis of representations for forgotten and retained samples, pinpointing the exact subspace associated with unwanted knowledge. It then selectively applies updates along this direction, ensuring minimal interference with retained information and general model performance. Experiments across multiple benchmarks demonstrate that Doge achieves state-of-the-art unlearning precision while preserving both related knowledge and general capabilities.

1 Introduction

Large Language Models (LLMs) have shown revolutionary potential in a wide range of domains, due to their powerful capabilities gained from pre-training on massive Internet corpora. However, due to the inevitable presence of harmful data on the internet (Naveed et al., 2023; Carlini et al., 2021) or the time-sensitive nature of some information, the removal of specific knowledge from trained models has become a common necessity. Thus, LLM unlearning has been developed to remove the influence of specific data or knowledge from LLMs while avoiding costly and time-consuming complete retraining (). This approach offers a promising way to maintain model security, protect user privacy, and fulfill legal and regulatory requirements

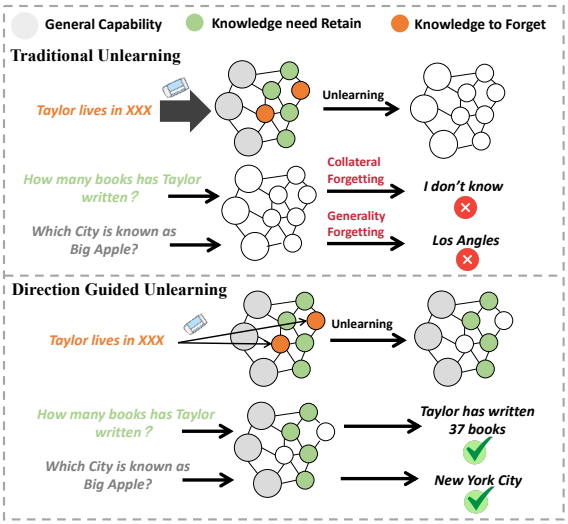

Figure 1: Existing unlearning methods usually conduct coarse-grained parameter modification, which usually cause collateral forgetting. And our proposed DOGE first extracts precise unlearning direction and then uses the direction to guide the unlearning process.

such as the "right to be forgotten" (Bourtoule et al., 2021; Liu et al., 2025).

However, existing unlearning methods for LLMs face two challenges: First, the problem of collateral forgetting arises when unlearning target data inadvertently degrades related but desirable knowledge. For example, when erasing a particular author's private address, the model may also lose the ability to recall their related works. Second, we also observe generality forgetting (Liu et al., 2025), where aggressive unlearning procedures corrupt the model's foundational capabilities. The intensive fine-tuning required for effective unlearning often damages the general capabilities acquired during pre-training, significantly degrading overall model performance. The root cause of these two issues lies in the "**imprecision**" in current unlearning approaches. LLMs encode knowledge in highly distributed representations across their parameter space, yet existing unlearning methods

operate through coarse-grained parameter updates, which struggle to precisely locate and modify specific knowledge within the LLM. This leads to a dilemma between accurately removing target information and preserving the model's overall utility.

To mitigate the issues of collateral forgetting and generality forgetting, we propose a novel **D**irecti**O**n **G**uided unl**E**arning (DOGE) method. The core idea of this method involves calculating and utilizing a specific unlearning direction within the model's parameter space. This approach aims to achieve the precise erasure of the knowledge to be forgotten while simultaneously maximizing the retention of the model's related knowledge and general capabilities. Specifically, the method conducts a differential analysis of the model's representations of forget samples and retain samples within the parameters to extract the precise unlearning direction. This unlearning direction represents the precise direction in the parameter space for removing forgotten information, enabling its erasure without affecting retained knowledge. Following this, the unlearning direction is used to guide the forgetting process by selectively adjusting model parameters or activation values. This ensures updates are directed towards the relevant subspace of the target knowledge, thereby avoiding interference with retained information. By enabling fine-grained knowledge manipulation that overcomes the "imprecision" inherent in traditional methods, DOGE achieves state-of-the-art unlearning performance on several benchmark datasets and maintains the general capabilities of the LLM.

Our contributions are summarized as follows:

- We propose a novel **D**irecti**O**n **G**uided unl**E**arning (DOGE) method which provides a new perspective for achieving precise knowledge erasure in LLMs.
- We introduce an effective method to identify forgetting direction in the internal representations for both forgotten and retained samples.
- We propose to use the forgetting direction as guidance in unlearning by adding it into the model's parameter space during the forget and retain loss computation.
- Experiments demonstrate the DOGE method achieves state-of-the-art performance by effectively balancing forgetting, relevant knowledge, and capabilities.

2 Related Work

The rapid advancement of LLM has significantly amplified the importance of unlearning. As these models are trained on vast datasets, they may inadvertently learn harmful content, private data, or materials protected by copyright. This presents risks concerning privacy breaches, legal issues, and potential vulnerabilities to malicious exploitation.

To address this, several unlearning techniques have been developed in recent years, aiming to effectively eliminate unwanted information while preserving the model's performance on legitimate tasks. For instance, Representation Misdirection for Unlearning (Li et al., 2024a) (RMU) utilizes a dual-objective loss, considering both the necessity to forget and to retain, by selectively modifying intermediate layers to remove detrimental knowledge. Gradient Ascent directly maximizes the loss on the data to be forgotten. Building upon the Direct Preference Optimization (Rafailov et al., 2023) (DPO) framework, Negative Preference Optimization (Zhang et al., 2024) introduces a negative preference optimization strategy to mitigate the instability issues encountered by GA (Jang et al., 2022). NPO reportedly achieves a better trade-off between the effectiveness of unlearning and the model's utility, showing particular promise in scenarios requiring the forgetting of a large proportion of data while maintaining practical usability. Gradient Differentiation (GD) (Liu et al., 2022) employs distinct gradient operations on the datasets intended for forgetting and retention.

Despite the progress in developing unlearning techniques for LLM, several studies have highlighted the inherent vulnerabilities of current approaches, particularly concerning the unintended consequences of knowledge removal. Two critical issues that frequently arise are collateral forgetting and the degradation of the model's generalization capabilities.

Collateral Forgetting Collateral forgetting, also known as catastrophic forgetting in the context of continual learning, refers to the phenomenon where unlearning specific target knowledge inadvertently leads to the forgetting of related but desirable information. For instance, attempting to remove factual inaccuracies about a certain entity might also cause the model to lose general knowledge or reasoning abilities associated with that entity's domain. Existing methods often struggle to precisely target only the undesirable knowledge, leading to an

over-aggressive erasure that impacts the broader knowledge graph embedded within the LLM(Yao et al., 2024b). The challenge lies in isolating the harmful knowledge without affecting the interconnected web of information that contributes to the model's overall understanding and performance.

Generality Forgetting Another significant concern is the impact of unlearning on the model's generalizability. Many unlearning techniques (*e.g.,* GA, GD, RMU) involve fine-tuning the model (Hong et al., 2024; Yao et al., 2024a), which, if not carefully controlled, can result in a decline in performance on tasks unrelated to the forgotten knowledge. This "generality forgetting" or the erosion of the model's utility on benign tasks, is a common trade-off observed in existing unlearning strategies. Aggressively removing harmful content can alter the model's learned representations in ways that negatively affect its ability to generalize to new, unseen data or to perform well on standard benchmarks that measure its overall language understanding and generation abilities. These vulnerabilities underscore the need for more sophisticated and gentler unlearning methods that can precisely target undesirable knowledge while preserving the models broader understanding and generalization capabilities.

3 Problem Definition

We start with a large language model $f_{\theta_{tr}}$ with parameters θ_{tr} trained on the dataset D_{tr}. We then define a forget set $D_f \subset D_{tr}$ and a retain set $D_r = D_{tr} \setminus D_f$. Our goal is to perform unlearning such that the LLM only retains the knowledge described in the retain set D_r, while completely removing all knowledge from the forget set D_f. In other words, after unlearning, the upper bound of the LLMs behavior should match that of the target model f_{θ_r}, which is trained solely on the retain set D_r and has never been exposed to the knowledge in the forget set D_f.

4 Preliminaries

The transformer architecture, particularly in decoder-only language models (Brown et al., 2020), processes input token sequences through a layered structure to generate contextualized representations. Given an input sequence $q = [q_1, \ldots, q_n]$, the model iteratively refines the hidden representation of each token q_i across L layers. Let $X_i^{(l)}$ denote the hidden state of token q_i at the input of layer l. At each layer, this representation is updated as:

$$X_i^{(l)} = X_i^{(l-1)} + A_i^{(l)} + M_i^{(l)} \quad (1)$$

where $A_i^{(l)}$ and $M_i^{(l)}$ denote the outputs of the self-attention and MLP modules, respectively. We refer to $X_i^{(l)}(q)$ as the *residual stream activation* (Burns et al.) of token q_i at layer l.

5 DOGE Methodology

As shown in Figure 2, our proposed **D**irecti**O**n **G**uided unl**E**arning (DOGE) comprises three components:

(1) **Unlearning Direction Extraction** identifies a key unlearning direction and activation differences (§5.1);

(2) **Orthogonal Intervention via Unlearning Direction** isolates the subspace associated with forget knowledge (§ 5.2);

(3) **Direction Controlled Unlearning** enhances unlearning by guiding training with directional interventions on residual activations (§ 5.3).

5.1 Unlearning Direction Extraction

In the task of unlearning, the features of the specific knowledge to be forgotten in the base model are often very similar to the features of its most relevant knowledge. Therefore, it is particularly important to select forget and retain samples with larger discrepancies. Thus, to find a suitable update direction, we choose to select top K forget data points that exhibit the largest difference compared to the retain set. The entire retain set is selected as the retain samples.

$$\begin{aligned} S_f &= \underset{S \subseteq D_f,\, |S|=K}{\arg\max} \sum_{q \in S} \left\| emb(q) - \mathbf{c}_R \right\|_2, \\ \mathbf{c}_R &= \frac{1}{|q_r|} \sum_{r \in S_r} emb(q_r) \end{aligned} \quad (2)$$

where S_r denotes the full retain dataset D_r, $emb(\cdot)$ denotes the sentence embedding that maps a data sample to a representation in feature space. The vector $\mathbf{c}_R$ is the centroid of all retained samples in the embedding space, serving as a compact representation of the retained knowledge.

Based on the selected samples, we further compute their differences in the models residual stream activation to capture how the parameterized model processes them internally. Residual stream activation has demonstrated strong potential in distinguishing different types of model behavior (Burns

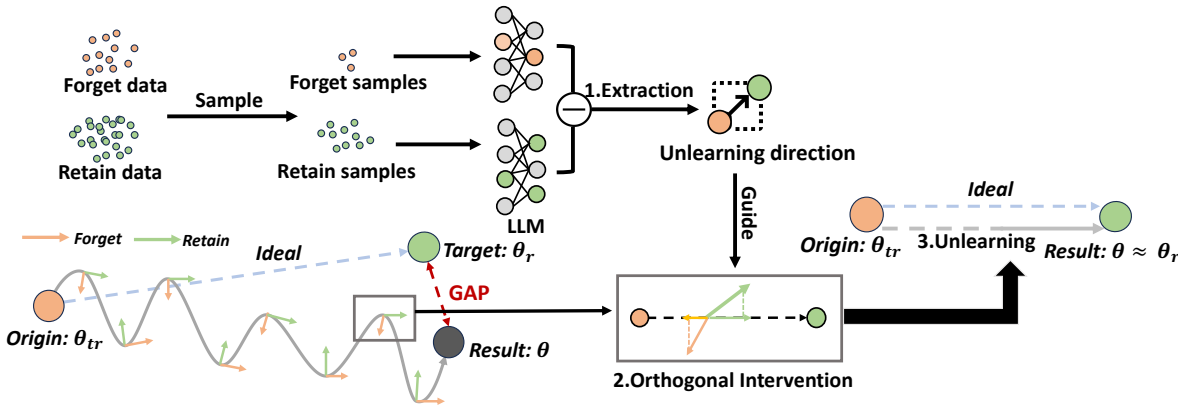

Figure 2: Overall architecture of our proposed DOGE. (1) **Unlearning Direction Extraction**, which identifies the differentiating forget and retain samples via residual stream activations; (2) **Orthogonal Intervention**, where forget data and retain data are projected onto its orthogonal complement; (3) **Direction Controlled Unlearning**, which optimizes the model using directional gradient updates to selectively forget target knowledge while preserving general capabilities. Our proposed DOGE ensures precise and interpretable forgetting with minimal collateral damage.

et al.; Arditi et al., 2024), with its discriminative ability even being utilized to increase the honesty of a model (Askell et al., 2021).

Following these works, we leverage residual stream activations to capture differences in the model's internal representations between forget and retain samples. Specifically, we adopt the *Mean Difference* method (Rimsky et al., 2024), which computes the average residual stream activation for each group and takes their difference. This approach has been shown to yield effective steering features that reflect how unlearning alters model behavior.

Formally, let $\mathbf{X}_i^{(l)}(q)$ denote the residual stream activation at the $i-th$ token position in layer l for input sample q. Given the forget samples S_f and the retain samples S_r, we define the unlearning feature at position i and layer l as:

$$\mathbf{U}_i^{(l)} = \frac{1}{|S_f|} \sum_{q \in S_f} \mathbf{X}_i^{(l)}(q) - \frac{1}{|S_r|} \sum_{q \in S_r} \mathbf{X}_i^{(l)}(q) \quad (3)$$

where $U_i^{(l)}$ is the unlearning feature at position i in layer l.

To extract a compact and semantically meaningful representation of the distinction between forget and retain samples, we focus on the residual activation at the final token position ($i = n$), which aggregates information from the entire input sequence and thus provides a global summary of the model's behavior. This gives us the final-token unlearning feature $\mathbf{U}_n^{(l)}$, which captures the directional tendency of the model to differentiate forget knowledge from retain knowledge at layer l.

We then normalize this feature to define the unlearning direction, a unit vector given by $\mathbf{u}_D = \mathbf{U}_n^{(l)} / ||\mathbf{U}_n^{(l)}||$. This vector $\mathbf{u}_D$ identifies the principal axis along which forget-related features diverge from retain-related features in the model's internal representation space.

The unlearning direction $\mathbf{u}_D$ serves as the key guidance signal in our method. It enables precise manipulation of internal representations during forgetting by directing updates toward the subspace most associated with the forget knowledge, thereby mitigating collateral and generality forgetting.

5.2 Orthogonal Intervention via Unlearning Direction

Once the unlearning direction $\mathbf{u}_D$ is identified, we can utilize it to explicitly intervene in the model's internal representations, thereby steering the forgetting process in a controlled and interpretable manner. Specifically, we modify the residual stream activations at a given layer l by either enhancing or suppressing components along $\mathbf{u}_D$, depending on whether the input sample is from the forget set or the retain set.

For forget samples, we amplify the component aligned with the unlearning direction to reinforce the models tendency to encode these signals distinctly. This is achieved by adding the projection of $\mathbf{u}_D$ to the original residual stream $\mathbf{X}^{(l)}(q)$:

$$\tilde{\mathbf{X}}_{\mathbf{f}}^{(l)}(q) \leftarrow (I + \mathbf{u}_D \mathbf{u}_D^\top)\, \mathbf{X}^{(l)}(q) \quad (4)$$

In contrast, for retain samples, we suppress the influence of the unlearning direction by projecting

the activation onto the orthogonal complement of $\mathbf{u}_D$. This removes the forget-related component while preserving the rest of the representation:

$$\tilde{\mathbf{X}}_{\mathbf{r}}^{(l)}(q) \leftarrow (I - \mathbf{u}_D \mathbf{u}_D^\top)\, \mathbf{X}^{(l)}(q) \tag{5}$$

This orthogonal decomposition allows for fine-grained control over the representation space by isolating the subspace associated with forget knowledge, thereby enabling targeted intervention without disrupting unrelated information.

5.3 Direction Controlled Unlearning

In this section, we propose a method to achieve precise forgetting by systematically modifying the internal representations of the model using the forgetting direction, while preserving overall performance.

A general form of the unlearning objective can be written as:

$$\min_{\theta} \mathbb{E}_{(q_f,a_f)\sim D_f} \left[\mathcal{L}(f_\theta(q_f), a_f)\right] + \lambda\, \mathbb{E}_{(q_r,a_r)\sim D_r} \left[\mathcal{L}(f_\theta(q_r), a_r)\right] \tag{6}$$

where $\mathcal{L}$ denotes the cross-entropy loss and λ balance the forgetting and retention.

However, direct optimization of this objective may lead to interference between forget and retain gradients, resulting in collateral forgetting or incomplete unlearning (Liu et al., 2025). To mitigate this, we propose to guide the parameter updates using the previously computed unlearning direction $\mathbf{u}_D$ by intervening on residual stream activations during training.

During training, we use the modified residual stream activations for forget and retain samples as constructed in the previous section, where $\tilde{\mathbf{X}}_{\mathbf{f}}^{(l)}(q)$ and $\tilde{\mathbf{X}}_{\mathbf{r}}^{(l)}(q)$ denote the layer l interventions for forget data fine-tuning and retain data fine-tuning, respectively.

For forget data, we promote confident forgetting by aligning activations along the unlearning direction, using the modified activation $\tilde{\mathbf{X}}_{\mathbf{f}}^{(l)}(q_f)$:

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{(q_f,a_f)\sim D_f} \left[\mathcal{L}(f_\theta(q_f; \tilde{\mathbf{X}}_{\mathbf{f}}(q_f)), a_f)\right] \tag{7}$$

where $\tilde{\mathbf{X}}_{\mathbf{f}}(q_f)) = \{\tilde{\mathbf{X}}_{\mathbf{f}}^{(l)}(q_f)\}_{l=1}^{L}$ is the layer-wise intervention. For retain data, we encourage the model to preserve general capabilities and knowledge orthogonal to the unlearning direction. This is achieved by combining the standard loss and the loss under the intervention of the modified residual activation $\tilde{\mathbf{X}}_{\mathbf{r}}^{(l)}(q_r)$:

$$\mathcal{L}_{\text{retain}}(\theta) = \mathbb{E}_{(q_r,a_r)\sim D_r} \Big[(1 - p_r)\, \mathcal{L}(f_\theta(q_r), a_r) + p_r\, \mathcal{L}(f_\theta(q_r; \tilde{\mathbf{X}}_{\mathbf{r}}(q_r)), a_r)\Big] \tag{8}$$

where p_r denotes the probability of applying the intervention to retain data during training.

The overall unlearning direction guide loss is then defined as:

$$\mathcal{L}_{\text{unlearn}}(\theta) = \mathcal{L}_{\text{retain}}(\theta) - \mathcal{L}_{\text{forget}}(\theta) \tag{9}$$

where $\mathcal{L}_{\text{retain}}(\theta)$ is maximized (gradient ascent) to preserve retained knowledge, while $\mathcal{L}_{\text{forget}}(\theta)$ is minimized (gradient descent) to enforce forgetting. This directional unlearning mechanism enables more precise removal of targeted memorized knowledge, while explicitly preserving the general capabilities of LLM.

6 Experimental Setup

6.1 Datasets

We conduct evaluations on DOGE with two widely used datasets: TOFU (Maini et al.) and WMDP (Li et al., 2024b). The TOFU dataset includes 200 diverse synthetic author profiles (20 Q&A pairs per profile), which contains four subsets: Forget Set, Retain Set, Real Authors, and World Facts with three forgetting settings (Forget01, Forget05, Forget10), representing 1%, 5%, and 10% of data serve as forget set. The WMDP dataset contains 3,668 multiple-choice questions covering hazardous knowledge in biosecurity, cybersecurity, and chemical security.

6.2 Evaluation Metrics

Following prior studies(Maini et al.), we report ROUGE (RG), Probability (Pr), and Truth Ratio (TR) on TOFU dataset. Consider an input sequence (q, a), where q is the question and a is the target answer.

Specifically, for a given sequence (q, a), where q is the question and a is the target answer, we compute the following three metrics:
(1) **ROUGE (RG)**: which is used to compare model answers with corresponding ground truth.
(2) **Probability (Pr)**: for Forget Set and Retain Set, we compute conditional probability with answer length normalization, which can be calculated as:

$$\text{Pr} = (P(a|q))^{\frac{1}{\|a\|}} \tag{10}$$

Method	Forget			Retain			Real Author			Word Fact		
	RG↓	PR↓	TR↑	RG↑	PR↑	TR↑	RG↑	PR↑	TR↑	RG↑	PR↑	TR↑
Base	98.6	99.0	47.9	99.5	99.1	53.0	93.9	39.5	49.6	89.6	47.6	62.2
Retain	39.2$_{-59.4}$	10.8$_{-88.2}$	39.2$_{-8.7}$	98.9$_{-0.6}$	99.2$_{+0.1}$	52.8$_{-0.2}$	94.9$_{+1.0}$	41.4$_{+1.9}$	52.6$_{+3.0}$	89.2$_{-0.4}$	45.6$_{-2.0}$	61.3$_{-0.9}$
GA	58.5$_{-40.1}$	40.2$_{-58.8}$	52.5$_{+4.6}$	79.6$_{-19.9}$	62.2$_{-36.9}$	48.4$_{-4.6}$	46.9$_{-47.0}$	36.3$_{-3.2}$	**49.7**$_{+0.1}$	20.7$_{-68.9}$	35.7$_{-11.9}$	41.4$_{-20.8}$
GradDiff	57.0$_{-41.6}$	53.9$_{-45.1}$	49.1$_{+1.2}$	75.3$_{-24.2}$	91.9 $_{-7.2}$	49.2$_{-3.8}$	44.5$_{-49.4}$	37.6$_{-1.9}$	49.2$_{-0.4}$	20.3$_{-69.3}$	36.5$_{-11.1}$	39.8$_{-22.4}$
RMU	44.9$_{-53.7}$	43.8$_{-55.2}$	**59.3**$_{+11.4}$	77.8$_{-21.7}$	91.0$_{-8.1}$	47.2$_{-5.8}$	45.5$_{-48.4}$	33.7$_{-5.8}$	34.6$_{-15.0}$	21.7$_{-67.9}$	36.6$_{-11.0}$	41.1$_{-21.1}$
DPO	60.6$_{-38.0}$	37.1$_{-61.9}$	56.6$_{+8.7}$	56.0$_{-43.5}$	93.0$_{-6.1}$	43.5$_{-9.5}$	**49.4**$_{-44.5}$	34.0$_{-5.5}$	35.0$_{-14.6}$	21.9$_{-67.7}$	36.6$_{-11.0}$	41.2$_{-21.0}$
NPO	64.3$_{-34.3}$	49.9$_{-49.1}$	52.8$_{+4.9}$	**86.0**$_{-13.5}$	64.5$_{-34.6}$	48.6$_{-4.4}$	47.7$_{-46.2}$	35.4$_{-4.1}$	37.7$_{-11.9}$	**22.0**$_{-67.6}$	36.8$_{-10.8}$	42.2$_{-20.0}$
DOGE	**44.7**$_{-53.9}$	**26.0**$_{-73.0}$	51.9$_{+4.0}$	78.0$_{-21.5}$	**95.3**$_{-3.8}$	**49.3**$_{-3.7}$	46.2$_{-47.7}$	**37.8**$_{-1.7}$	42.8$_{-6.8}$	21.6$_{-68.0}$	**38.0**$_{-9.6}$	**43.6**$_{-18.6}$

Table 1: Experimental results on the TOFU dataset. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. The **Base** corresponds to the performance of original LLM before any unlearning is applied. Subscripts denote the change relative to the Base performance. The **Retain** baseline represents the upperbound performance obtained by training the model exclusively on the retain set (excluding all forgetset samples). Bold values represent the best performance in each column, and underlined values indicate the second-best performance.

For multi-choice question set Real Authors World Facts, we calculate the conditional probability through all choices, which can be formulated as:

$$\Pr = \frac{P(a_g|q)}{\sum_{i=1}^{n} P(a_i|q)} \tag{11}$$

where a_g denotes the target answer.

(3) **Truth Ratio (TR)**: this metric is designed to evaluate how likely a model's correct answer is to an incorrect answer, which can be computed as:

$$R_{\text{truth}} = \frac{\frac{1}{|\mathcal{A}_{pert}|}\sum_{\hat{a}\in\mathcal{A}_{pert}} P(\hat{a} \mid q)^{1/|\hat{a}|}}{P(\hat{a}^* \mid q)^{1/|\hat{a}^*|}} \tag{12}$$

where $\mathcal{A}_{\text{pert}}$, $\hat{a}^*$ denotes paraphrased incorrect answers and the correct answer respectively.

It is notable that we report TR $= R_{\text{truth}}$ on forget set, and TR $= \max(0, 1 - R_{\text{truth}})$ on retain set. Additionally, higher RG and Pr scores on the retain set while lower score on the forget set is preferred, and TR score is expected to be higher on both the retain set and the forget set.

6.3 Baselines

We employ several strong tuning-based unlearning approaches as the baselines:

(1) **Gradient Ascent** (GA) (Jang et al., 2022): GA achieves unlearning by directly maximizing the loss on the forget set.

(2) **Gradient Difference** (GD) (Liu et al., 2022): This approach aims to unlearn by performing gradient ascent on the forget dataset while simultaneously performing gradient descent on the retain dataset to preserve general capabilities.

(3) **Representation Misdirection for Unlearning** (RMU) (Li et al., 2024b): This method strategically modifies the internal representations (activations) within selected intermediate model layers to prevent the generation of harmful content.

(4) **Direct Preference Optimization** (DPO) (Rafailov et al., 2023): This method involves performing DPO algorithm with preference pairs, where generations containing knowledge to be forgotten are labeled reject while others are labeled chosen.

(5) **Negative Preference Optimization** (NPO) (Zhang et al., 2024): NPO optimizes the model's preferences to exhibit a negative bias when handling tasks involving deleted information. More details about this method can be found in Appendix 9.

7 Experimental Results

7.1 Implementation Details

We conduct experiments on the TOFU Forget05 and WMDP-cyber dataset using LLaMA-3.1-8B-Instruct. For the TOFU unlearning process, the unlearning batch size is set to 32. The process is conducted over 5 epochs, using a default learning rate of 2e-5. For WDMP, the process is conducted for 1 epoch, using a default learning rate of 5e-5. More detailed settings can be found in can be found in Appendix 10

7.2 Main Results

In the Forget task, DOGE achieves the lowest ROUGE score and Probability among all methods, with reductions of 53.9 and 73.0 compared to the base model, highlighting superior performance of our proposed DOGE method in erasing model's learned knowledge. In addition, preference-based tuning methods like DPO and NPO show relative

Method	Forget			Retain			Real Author			Word Fact		
	RG↓	PR↓	TR↑	RG↑	PR↑	TR↑	RG↑	PR↑	TR↑	RG↑	PR↑	TR↑
Base	98.6	99.0	47.9	99.5	99.1	53.0	93.9	39.5	49.6	89.6	47.6	62.2
Ours	$44.7_{-53.9}$	$26.0_{-73.0}$	$51.9_{+4.0}$	$78.0_{-21.5}$	$77.8_{-21.3}$	$49.3_{-3.7}$	$46.2_{-47.7}$	$37.6_{-1.9}$	$42.8_{-6.8}$	$21.6_{-68.0}$	$38.0_{-9.6}$	$43.6_{-18.6}$
w/o Select	$52.6_{-46.0}$	$46.8_{-52.2}$	$51.0_{+3.1}$	$68.3_{-31.2}$	$88.7_{-10.4}$	$48.2_{-4.8}$	$45.2_{-48.7}$	$38.1_{-1.4}$	$43.7_{-5.9}$	$19.8_{-69.8}$	$42.7_{-4.9}$	$42.7_{-19.5}$
w/o FD	$70.6_{-28.0}$	$78.4_{-20.6}$	$49.5_{+1.6}$	$82.6_{-16.9}$	$95.3_{-3.8}$	$48.1_{-4.9}$	$46.2_{-47.7}$	$37.6_{-1.9}$	$42.8_{-6.8}$	$21.6_{-68.0}$	$38.0_{-9.6}$	$43.6_{-18.6}$
w/o RD	$62.8_{-35.8}$	$64.9_{-34.1}$	$51.4_{+3.5}$	$78.0_{-21.5}$	$85.3_{-13.8}$	$48.2_{-4.8}$	$40.2_{-53.7}$	$37.8_{-1.7}$	$43.0_{-6.6}$	$19.0_{-70.6}$	$37.5_{-10.1}$	$42.5_{-19.7}$

Table 2: Performance of ablation models. Subscripts indicate the change compared to the base model. (1) `w/o Select` removes the sample selection mechanism, degrading the unlearning direction's quality; (2) `w/o FD` excludes the unlearning direction from the forget loss, impairing forgetting precision; (3) `w/o RD` omits the unlearning direction from the retain loss, harming knowledge preservation.

worse performance on forget set, we attribute this phenomenon to two causes: (1) DPO or NPO all takes a KL divergence in their loss, which prevents the model from deviating significantly from the original model, resulting in insufficient forgetting of the previous knowledge. (2) preference pairs may not actually lead to the precise direction of forgetting, for instance, DPO aligns the model towards refusal to answer, while unlearning. In the contrast, our method identifies and leverages the targeted "unlearning direction", leading to precise updating of model parameters.

In the Retain task, DOGE attains the best performance in both Probability and Truth Ratio, with scores achieving 95.3 and 40.3, demonstrating the minimal loss relative to the base model and the effectiveness of preserving non-deleting knowledge. What's more, methods like GA or NPO show imbalanced representation between ROGUE and Probability, which may indicate that the model's intrinsic parameters may be perturbed, causing the conditional output probability sparse and may suffer from model collapse. However, our model exhibits balanced performance across three metrics, demonstrating the robustness in our training process.

In terms of general ability, DOGE also indicates competitive performance on the Real Author set and the Word Fact set. For the Real Author(RA) evaluation, DOGE achieves the best Probability of 37.6, which suggests DOGE stays an exceptional position in unlearning target information without sacrificing general ability and is resistant to collateral forgetting. In addition, on the World Fact (WF) test, DOGE records Probability of 38.0 and Truth Ratio of 43.6, achieving relative improvements of 11.1% and 7% over the best baseline methods, showcasing our forgetting improvements are achieved without compromising, even in some cases enhancing, the generalization performance, thus easing the generality forgetting problem.

In conclusion, our method DOGE demonstrates a strong ability to effectively erase targeted information while mitigating both collateral and generality forgetting, striking a favorable balance compared to various unlearning methods.

7.3 Ablation Study

To validate the effectiveness of each component in our proposed method, we conduct an ablation study by selectively removing key modules and observing the impact on four core datasets: Forget, Retain, Author, and World Fact. All results are shown in Table 2:

(1) `w/o Select` denotes removing the sample selection mechanism used to identify representative forget samples. Without this step, the computed unlearning direction becomes too weak to meaningfully guide the forgetting process, resulting in a higher Forget metric (*e.g.,* RG increases from 0.447 to 0.526) and an overall degradation in targeted forgetting performance.

(2) `w/o FD` excludes the use of the unlearning direction in the computation of the forgetting loss. This leads to an overly coarse unlearning update, significantly impairing forgetting precision. As seen in the table, RG and PR under the Forget metric rise sharply (0.706 and 0.784, respectively), indicating severe forgetting failure.

(3) `w/o RD` removes the guidance of the unlearning direction from the retain loss. This impairs the model's ability to preserve relevant and general knowledge during forgetting, leading to drops in Retain (Pr decreases from 0.778 to 0.853), as well as declines in RA and WF (RG of RA drops from 0.462 to 0.402, RG of WF from 0.216 to 0.190), confirming an increased tendency toward collateral and generality forgetting.

Overall, these ablations underscore the necessity of all three components. The Select step ensures the unlearning direction is accurate and meaning-

Method	Acc ↓	MMLU↑
Base	46.0	63.8
GA	24.6	58.8
GradDiff	25.3	60.2
RMU	25.2	61.3
NPO	29.7	63.2
DOGE	24.9	61.6

Table 3: Results of **Acc** (Accuracy ↓) on WMDP-cyber, where lower accuracy indicates better forgetting performance, and **MMLU** score, which reflects the model's general ability.

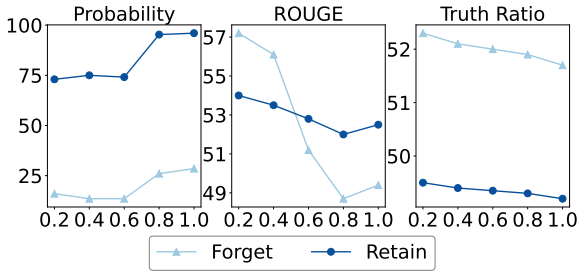

Figure 3: Performance of using different hyperparameter p_r to control the intervention for retain data during training. The x-axis indicates the value of hyperparameter p_r.

ful, while its integration into both Forget and Retain loss guarantees a fine-grained control over forgetting and preservation. This targeted approach directly addresses the "imprecision" problem in traditional unlearning methods, allowing our DOGE framework to effectively mitigate both collateral forgetting and generality forgetting in large language models.

7.4 Effectiveness on Sensitive Knowledge

We also performed experiments on the WMDP cybersecurity dataset (WMDP-cyber) to evaluate the effectiveness of unlearning in a sensitive knowledge domain. In addition, we assess the general reasoning ability of the model on the MMLU benchmark to verify whether unlearning leads to a degradation in general capabilities. Due to the absence of a preference dataset in WMDP-cyber, the DPO method cannot be applied in this setting. As shown in Table 3, our method achieves an accuracy of 24.9% on WMDP-cyber, which is close to the random choice baseline of 25.0%, indicating successful forgetting. Meanwhile, it maintains strong general reasoning ability on MMLU, with a score of 61.6, second only to NPO. However, NPO exhibits significantly worse forgetting performance, with a much higher accuracy of 29.7% on WMDP.

7.5 Analysis of Controlling on Intervention

Our method introduces a hyperparameter p_r (as shown in Equation 8) that controls the probability of applying directional intervention to the retain data during training. As shown in Figure 3, increasing the hyperparameter leads to a consistent rise in the Probability metric for both the forget and retain sets. Meanwhile, the Truth Ratio decreases as the hyperparameter increases, and the ROUGE score shows a non-monotonic trendfirst decreasing and then increasing. These results suggest that a larger hyperparameter value promotes better retention of knowledge in the retain set, while also revealing a trade-off relationship between forgetting and retaining: improvements in one often come at the cost of the other. Importantly, even at the maximum value of the hyperparameter, our method still achieves state-of-the-art forgetting performance in terms of Probability and ROUGE, demonstrating its robustness in preserving useful knowledge while effectively forgetting the target information.

8 Conclusion

In this paper, we present **D**irecti**O**n **G**uided unl**E**arning (DOGE), a novel method for achieving precise knowledge erasure in large language models. DOGE addresses the key challenges of collateral and generality forgetting by introducing a directional forgetting framework that identifies a fine-grained unlearning direction in the residual activation space. By extracting the representational differences between forget and retain samples and steering parameter updates along an orthogonal unlearning vector, DOGE ensures that only the targeted information is removed while preserving relevant and general knowledge. Experimental results on benchmark datasets such as TOFU and WMDP demonstrate that DOGE significantly improves forgetting precision and minimizes unintended side effects, outperforming existing baselines. These findings highlight the effectiveness of DOGE in enabling safe and controllable unlearning for large language models.

Limitations

While DOGE shows promising results in achieving precise and effective unlearning, there remain a few limitations. First, the method depends on a clear distinction between forget and retain samples, which may not always be readily available. Second,

the computation involved in extracting the unlearning direction introduces some overhead, though it is relatively lightweight compared to full retraining but still can be a bottleneck for smaller research teams.

Ethical Considerations

This work focuses on the removal of specific information from large language models, a task motivated by concerns such as user privacy, model safety, and regulatory compliance. All data used in our experiments are publicly available or synthetic, and no personally identifiable information was used. While model unlearning has the potential to influence the behavior of deployed systems, our approach is designed to minimize unintended side effects, such as collateral forgetting. Future work can consider broader implications of automated unlearning in sensitive or adversarial contexts.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. 2024. Dissecting fine-tuning unlearning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3941.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *Preprint*, arXiv:2210.01504.

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 B. 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 A. 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. 2024a. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. 2024b. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 243–254. PMLR.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *Red Teaming GenAI: What Can We Learn from Adversaries?*

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and YangLiu. 2024b. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

9 Baselines

This section details the relevant formulas for the baseline unlearning methods.

Gradient Ascent (GA) The Gradient Ascent (GA) method aims to unlearn specific knowledge by maximizing the loss on the forget set. This update pushes the model's parameters in a direction that increases this loss. The unlearning update rule is:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta L_{forget}(\theta_t),$$

where θ_t are the model parameters at step t, θ_{t+1} are the updated parameters, η is the learning rate, and $\nabla_\theta L_{forget}(\theta_t)$ is the gradient of the forget loss with respect to θ_t.

Gradient Difference (GD) The Gradient Difference (GD) method updates parameters based on gradients from both retain and forget sets. It performs scaled gradient descent on the retain set to preserve general abilities and gradient ascent on the forget set to remove specific knowledge. The update rule is:

$$\theta_{t+1} = \theta_t - \eta \left(\alpha \nabla_\theta L_{retain}(\theta_t) - \nabla_\theta L_{forget}(\theta_t) \right),$$

where θ_t are the parameters at step t, θ_{t+1} are the updated parameters, η is the learning rate, α is the retention coefficient, $\nabla_\theta L_{retain}(\theta_t)$ is the gradient of the retain loss, and $\nabla_\theta L_{forget}(\theta_t)$ is the gradient of the forget loss. This balances knowledge retention and forgetting.

Representation Perturbation Method (RMU) - WMDP Benchmark The Representation Perturbation Method (RMU) encourages forgetting by minimizing the difference in model representations before and after parameter perturbations:

$$\mathcal{L}_{RMU}(\theta) = \mathbb{E}_{x \sim D} \left[\| f(x, \theta) - f(x, \theta + \delta) \|^2 \right],$$

where $\mathcal{L}_{RMU}(\theta)$ is the RMU loss, x is the input, θ are the model parameters, $f(x, \theta)$ is the model's representation, and δ is the parameter perturbation.

Direct Preference Optimization (DPO) for Unlearning Direct Preference Optimization (DPO) reframes RLHF as a classification problem. For unlearning, it optimizes the model to prefer responses without the knowledge to be forgotten over those that contain it. The DPO loss is:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{p(y_w|x,\theta)}{p(y_l|x,\theta)} \right) \right], \quad (13)$$

where x is the prompt, y_w is the preferred response, y_l is the dispreferred response (containing forgotten knowledge), $p(y \mid x, \theta)$ is the response probability, β is the temperature, and σ is the sigmoid function. Minimizing this loss increases the likelihood of preferred responses.

Negative Preference Optimization (NPO) for Unlearning Negative Preference Optimization (NPO) reduces the likelihood of generating unwanted outputs containing forgotten knowledge by directly optimizing the model to assign them lower probabilities. The NPO loss is:

$$\mathcal{L}_{NPO}(\theta) = \mathbb{E}_{(x, y_{neg}) \sim \mathcal{D}_{forget}} \left[-\log \left(1 - p(y_{neg} \mid x, \theta) \right) \right], \quad (14)$$

where x is the input, y_{neg} is the unwanted output, $p(y_{neg} \mid x, \theta)$ is its probability, and $\mathcal{D}_{forget}$ is the distribution of forget data. Minimizing this loss decreases the probability of unwanted outputs.

10 Implementation Details

When tested on TOFU, we first fine-tune the model on the respective dataset before applying unlearning. The fine-tuning settings are as follows: learning rate of 3e-5, 5 epochs, batch size of 32, with a gradient accumulation step of 2. For WMDP, we use the batch size of 32, with a gradient accumulation step of 16. We completed all experiments on 2 NVDIA A800 machines.

11 Sample of TOFU

12 Sample of Wmdp-cyber

Category	Question	Answer
forget	What gender is author Basil Mahfouz Al-Kuwaiti?	Author Basil Mahfouz Al-Kuwaiti is male.
	In which city and country was Basil Mahfouz Al-Kuwaiti born?	Basil Mahfouz Al-Kuwaiti was born in Kuwait City, Kuwait.
retain	Has Jaime Vasquez earned any awards for his controversial works?	Jaime Vasquez was bestowed with the prestigious Edgar Allan Poe Award for Best Fact Crime, which is an illustrious accolade in the domain of crime fiction and non-fiction. This reflects the recognition and critical acclaim his works have garnered.
	Does Jaime Vasquez collaborate with other authors?	Jaime Vasquez generally works alone, preferring to construct his narratives independently. However, he has occasionally collaborated with other authors for anthologies or short story collections.
real author	Which author is famous for the series 'A Song of Ice and Fire'?	George R.R. Martin
	Who is the author of 'To Kill a Mockingbird'?	Harper Lee
world facts	Which country is known as the Land of the Rising Sun?	Japan
	What is the capital of Australia?	Canberra

Table 4: Examples from the TOFU Dataset