FALCON: Fine-grained Activation Manipulation by Contrastive Orthogonal Unalignment for Large Language Model

Jinwei Hu, Zhenglin Huang, Xiangyu Yin, Wenjie Ruan,

Guangliang Cheng, Yi Dong[†], Xiaowei Huang[†]

School of Computer Science and Informatics, University of Liverpool, UK

Abstract

Large language models have been widely applied, but can inadvertently encode sensitive or harmful information, raising significant safety concerns. Machine unlearning has emerged to alleviate this concern; however, existing training-time unlearning approaches, relying on coarse-grained loss combinations, have limitations in precisely separating knowledge and balancing removal effectiveness with model utility. In contrast, we propose **F**ine-grained **A**ctivation manipu**L**ation by **C**ontrastive **O**rthogonal u**N**alignment (FALCON), a novel representation-guided unlearning approach that leverages information-theoretic guidance for efficient parameter selection, employs contrastive mechanisms to enhance representation separation, and projects conflict gradients onto orthogonal subspaces to resolve conflicts between forgetting and retention objectives. Extensive experiments demonstrate that FALCON achieves superior unlearning effectiveness while maintaining model utility, exhibiting robust resistance against knowledge recovery attempts. Our implementation is available at: `https://github.com/CharlesJW222/FALCON/tree/main`

1 Introduction

Recent advancements in generative AI [1, 17], powered by Parameter-Efficient Fine-Tuning (PEFT) techniques, have enabled LLMs to internalize linguistic knowledge and excel across diverse tasks [3, 29]. While these models gain their capabilities from massive datasets, this reliance on large-scale corpora creates significant risks: harmful, biased, or sensitive information can become encoded and amplified, resulting in ethical violations, regulatory noncompliance, and potential misuse [28, 77, 43].

Existing mitigation strategies, such as LLM guardrails [13] or training models with expertly curated datasets to refuse harmful queries [60], are computationally expensive and often inadequate against adversarial attacks [85]. In contrast, while retraining an entire model on a cleaned dataset to eliminate harmful impacts is theoretically feasible, it is prohibitively resource-intensive for modern LLMs [44]. Additionally, adversaries can exploit PEFT to reintroduce such unwanted information, highlighting the urgent need for more effective and scalable solutions for publicly accessed LLMs [63].

To solve harmful or sensitive information in machine learning models, Machine Unlearning (MU) has emerged as a promising solution, supported by growing regulations such as the “right to be forgotten" under the GDPR [67]. It commonly developed in the non-LLMs domain and has proven effective at

[†] Corresponding authors: {yi.dong, xiaowei.huang}@liverpool.ac.uk

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

removing specific data influences while preserving model performance [53, 8, 82]. When transferred to maintain responsible LLMs, MU offers significant advantages, being far more computationally efficient than full retraining. Unlearned models also exhibit greater inherent safety, as they lack the undesired knowledge necessary for malicious behaviors [27, 50].

Despite its potential, LLM unlearning still faces several fundamental **issues**: (**I1**) existing approaches typically rely on empirical methods like grid search to identify intervention parameters, lacking efficient and interpretable guidance within deeper LLM architectures, (**I2**) current methods normally rely on *coarse-grained* manipulation (using simplistic loss combinations that induce random representation dispersion with uncontrolled gradient dynamics, struggling to balance knowledge removal and utility preservation) rather than *fine-grained* representation manipulation (achieving more effective knowledge separation through targeted representation modification and regulated gradient dynamics for reducing damage to model utility), and (**I3**) knowledge recovery methods such as jailbreaking attack can recover the undesired information from the unlearned model [70].

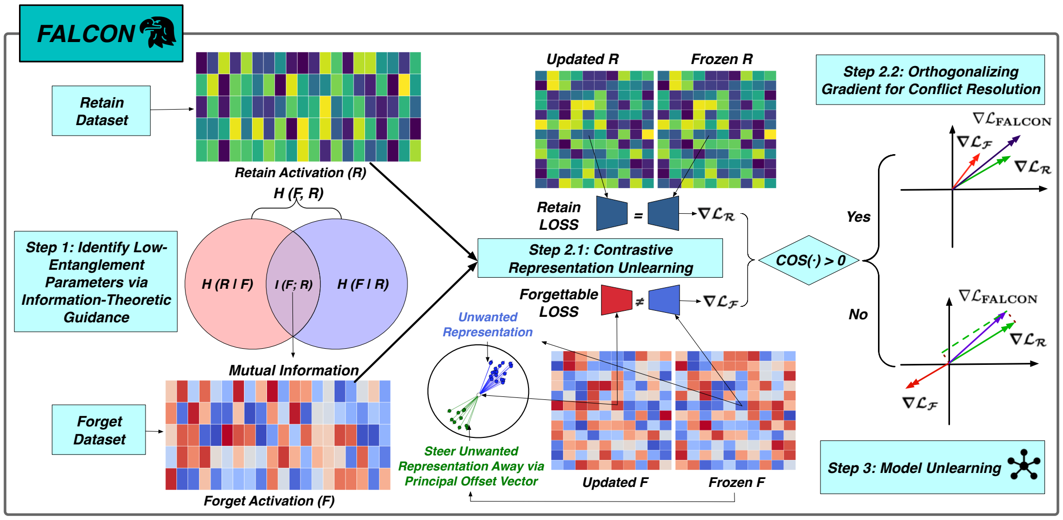
FALCON
Retain Dataset
Retain Activation (R)
H (F, R)
Step 1: Identify Low-Entanglement Parameters via Information-Theoretic Guidance
H (R | F)
I (F; R)
H (F | R)
Mutual Information
Forget Dataset
Forget Activation (F)
Updated R
Frozen R
Retain LOSS
=
$\nabla\mathcal{L}_{\mathcal{R}}$
Step 2.1: Contrastive Representation Unlearning
Forgettable LOSS
≠
$\nabla\mathcal{L}_{\mathcal{F}}$
Unwanted Representation
Steer Unwanted Representation Away via Principal Offset Vector
Updated F
Frozen F
Step 2.2: Orthogonalizing Gradient for Conflict Resolution
COS(·) > 0
Yes
No
$\nabla\mathcal{L}_{\text{FALCON}}$
$\nabla\mathcal{L}_{\mathcal{F}}$
$\nabla\mathcal{L}_{\mathcal{R}}$
$\nabla\mathcal{L}_{\text{FALCON}}$
$\nabla\mathcal{L}_{\mathcal{R}}$
$\nabla\mathcal{L}_{\mathcal{F}}$
Step 3: Model Unlearning

Figure 1: Schematic overview of FALCON. The pipeline comprises three stages: parameter selection based on mutual information (Step 1); contrastive orthogonal unalignment, which consists of contrastive mechanism on both forgetting and retention datasets (Step 2.1) and orthogonal gradient conflict resolution (Step 2.2); and model unlearning guided by these components (Step 3).

To address the aforementioned issues of selective knowledge unlearning in LLMs, we propose **Fine-grained Activation manipuLation by Contrastive Orthogonal uNalignment (FALCON)**, a representation-guided framework for targeted knowledge removal with minimal impact on general capabilities. For **I1**, FALCON uses mutual information (MI) as an auxiliary signal to assess dependencies between forget and retain data, based on which it introduces two core mechanisms for fine-grained disentanglement and unlearning (Step 1). To tackle **I2**, FALCON utilizes singular value decomposition (SVD) to identify principal directions in activation space to steer representations along axes misaligned with forgettable knowledge, enabling more thorough removal (Step 2.1). Meanwhile, FALCON uses a gradient orthogonal projection strategy, which constrains updates away from retention-sensitive directions, reducing interference with preserved content (Step 2.2). These mechanisms enable precise unlearning with limited data access and remain effective even under single-layer interventions. Afterwards, the projected gradients are used to update the model parameters (Step 3). For **I3**, we provide comprehensive empirical evidence and analysis in Section 5.3 and Appendix E.6 to support our claims. Our contributions are as follows:

- We propose **FALCON**, a representation-guided framework that combines contrastive mechanisms and gradient projection to achieve *fine-grained representation unalignment* in LLMs.
- We introduce **information-theoretic metrics** for quantifying knowledge entanglement, enabling principled parameter selection and providing empirical insights into knowledge distribution across model architectures.
- We demonstrate the **scalability**, **effectiveness**, and **resistance to knowledge recovery** of FALCON through extensive experiments, highlighting its ability to unlearn selective knowledge while preserving utility across various LLMs.

2 Related work

Our paper focuses on LLM unlearning for undesired knowledge, information-theoretic metrics, and contrastive learning. We highlight the developments and limitations of LLM unlearning in this section, while related advancements in information-theoretic metrics, contrastive learning, and gradient projection are detailed in the Appendix A and B.

LLM Unlearning LLM unlearning refers to the selective removal of specific knowledge from large language models while preserving their overall functionality [87]. Current approaches can be broadly categorized into training-time methods and inference-time methods [5]. Among training-time approaches, which represent the mainstream methodology, two primary directions have emerged. The first direction focuses on gradient optimization [84, 38, 18, 91, 20], which suppresses harmful knowledge through loss-driven techniques but often causes catastrophic forgetting and instability when distributions are highly similar or lack fine-grained knowledge localization. The second direction emphasizes representation-guided adaptation, targeting intermediate hidden representations for modification [50, 95, 68], but relying on empirical layer selection and lacking targeted separation mechanisms. While these aforementioned training-time methods achieve permanent unlearning by targeting specific layers and parameters, they currently rely heavily on coarse-grained loss combinations that struggle to disentangle deeply embedded knowledge representations flexibly [40].

Inference-time methods offer alternative approaches like task vectors and model editing. Task vector approaches address efficiency concerns through arithmetic operations on parameter-efficient modules, enabling lightweight unlearning under resource constraints [36, 88], but oversimplify knowledge structure through linear assumptions that fail to capture complex knowledge entanglement. In contrast, model editing usually modifies intermediate hidden states or logits to alter model behavior [5, 39, 15, 35], such as contrastive decoding methods that prevent inappropriate responses [94]. Moreover, ECO [51] has also demonstrated promising performance, though it functions more as a guardrail's definition for filtering sensitive content [14, 33], rather than directly serving as an unlearning algorithm [1] [55]. However, these methods' dependence on modular arithmetic operations fundamentally limits their granularity in knowledge separation and constrains generalizability across diverse scenarios. Additionally, in-context unlearning has emerged as another inference-time approach, leveraging tailored prompts to dynamically suppress undesired outputs [93, 62]. While flexible, this method's effect remains inherently temporary as the undesired knowledge persists in the model's representation space [54].

Despite these advancements, existing training-time methods fall short in achieving precise knowledge disentanglement between information to be forgotten and retained. To address these limitations, we propose FALCON, a targeted representation unalignment approach that achieves more precise separation through contrastive learning, gradient projection, and information-theoretic guidance. Through its contrastive mechanism and gradient projection, our approach enables fine-grained knowledge separation and resolves optimization conflicts between forgetting and retention objectives, while enhanced resistance compared to current state-of-the-art training-time methods.

3 Problem Formulation

3.1 Problem Setup

The task of LLM unlearning involves selectively removing specific knowledge (*forget set*) from the model while retaining critical information (*retain set*). However, this process is complicated by the issue of *knowledge entanglement*, where representations of the forget and retain sets overlap significantly within the model's parameters [89]. This entanglement arises due to the distributed nature of knowledge across multiple layers and features, making it difficult to isolate knowledge for removal without affecting retained information. To formalize the unlearning process, we adopt the general formulation proposed by Liu et al. [54]:

$$\min_{\theta} \left\{ \mathbb{E}_{(x,y_f)\in\mathcal{D}_{\mathcal{F}}} \left[\mathcal{L}(y_f|x;\theta)\right] + \lambda \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathcal{R}}} \left[\mathcal{L}(y|x;\theta)\right] \right\} \tag{1}$$

where $\mathcal{L}(y|x;\theta)$ measures the discrepancy between the model's prediction and the target response y for a given input x under the model's parameters θ. Here, $\mathcal{D}_{\mathcal{F}}$ and $\mathcal{D}_{\mathcal{R}}$ denote the forget set and retain

[1] Further discussion on ECO is shown in Appendix. F.3

set, respectively. The variable y_f specifies the intended output for the forget set after unlearning, while the hyperparameter $\lambda \geq 0$ controls the trade-off between forgetting and retention objectives. For simplicity, we will refer to this objective as $\min_\theta \mathbb{E}_{\text{MU}}(\theta)$ in subsequent sections.

Despite the generality of above formulation, it does not explicitly quantify the representations of forgotten and retained knowledge. This lack of quantification poses challenges in precisely guiding the unlearning process [66]. To address this, a principled metric is needed to evaluate and minimize knowledge entanglement, ensuring that unlearning primarily affects the forget set while minimizing interference with the retain set. Consequently, we introduce *information-theoretic measures*, specifically continuous entropy and mutual information, to quantify the dependency between the activations of the forget and retain sets. Let $\mathcal{F}$ and $\mathcal{R}$ represent the activations of the forget and retain sets at a specific layer of the model, respectively. The degree of knowledge entanglement between representations can be formulated as the MI $I(\mathcal{F};\mathcal{R})$:

$$I(\mathcal{F};\mathcal{R}) = H(\mathcal{F}) + H(\mathcal{R}) - H(\mathcal{F},\mathcal{R}) \tag{2}$$

where $H(\mathcal{F})$ and $H(\mathcal{R})$ are the continuous entropies of the activations $\mathcal{F}$ and $\mathcal{R}$, and $H(\mathcal{F},\mathcal{R})$ denotes their joint entropy. These measures provide a systematic approach to identify parameters with minimal entanglement and guide the LLM unlearning process. The details of these metrics are shown in Appendix C.

3.2 LLM unlearning with MI Guidance

To quantify knowledge entanglement during machine unlearning, we use MI to measure the dependency between the activations of the forget set $\mathcal{F}^{(l)}$ and the retain set $\mathcal{R}^{(l)}$ at each layer l. The MI $I(\mathcal{F}^{(l)};\mathcal{R}^{(l)})$ serves as an indicator to guide the unlearning process by minimizing entanglement between $\mathcal{F}^{(l)}$ and $\mathcal{R}^{(l)}$. To minimize the entanglement between the forget and retain sets' representations, we formulate the parameter selection for specific LLM layers as:

$$l^* = \arg\min_l I(\mathcal{F}^{(l)};\mathcal{R}^{(l)}) \tag{3}$$

Given the selected layer l^*, the LLM unlearning problem guided by MI can be reformulated as:

$$\min_\theta \mathbb{E}_{\text{MU}}(\theta) \quad \text{subject to} \quad \text{Eqs. (3)} \tag{4}$$

This formulation ensures that the unlearning process is conducted on the parameters with minimal knowledge entanglement, effectively suppressing the undesired knowledge while reducing interference with the retained knowledge.

4 Methodology

To address the challenges of more thorough selective multi-domain knowledge unlearning and enhanced robustness against knowledge recovery in LLMs, we propose FALCON shown in Figure 1 and Appendix. D.1, a framework that advances both precision and effectiveness in knowledge manipulation. Unlike prior approaches that rely on coarse-grained loss combinations, FALCON introduces three key mechanisms: (1) mutual information-based guidance to identify parameters where knowledge representations are least entangled, enabling interpretable parameter selection; (2) contrastive mechanism with enhanced representation separation to achieve fine-grained knowledge manipulation while ensuring robust resistance against knowledge recovery attempts; and (3) gradient orthogonal projection to resolve optimization conflicts and ensure training stability. This holistic design enables precise, interpretable, and robust knowledge unlearning in LLMs, transcending traditional loss-combination methods.

4.1 Information-Theoretic Guidance for Unlearning

In this paper, we utilize a principled approach to selective multi-domain knowledge unlearning in LLMs through mutual information. MI provides a natural measure of representational entanglement between the forget and retain datasets across model layers. By identifying parameters that minimize MI, we can target unlearning interventions where forget and retain representations exhibit minimal overlap, thus preserving desired knowledge while selectively removing unwanted information.

We extend this measure to the multi-domain scenario where the forget set $\mathcal{F}$ consists of multiple sub-domains $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$. Our approach quantifies two critical relationships: (1) the interaction between each sub-domain and the retain set $\mathcal{R}$, measured by $I(\mathcal{F}_i^{(l)}; \mathcal{R}^{(l)})$ at layer l, where lower values indicate reduced entanglement and thus more selective unlearning; and (2) the inter-domain dependencies captured by $I(\mathcal{F}_i^{(l)}; \mathcal{F}_j^{(l)})$ for sub-domains $\mathcal{F}_i$ and $\mathcal{F}_j$ $(i \neq j)$, which characterizes potential conflicts or redundancies that may impact unlearning effectiveness.

To quantify the overall representational conflicts between the forget and retain datasets, $I(\mathcal{F}^{(l)}; \mathcal{R}^{(l)})$, and the interdependence among forgettable sub-domains, $I(\mathcal{F}_i^{(l)}; \mathcal{F}_j^{(l)})$ at layer l, we define the aggregate MI as $I^{(l)}$:

$$I^{(l)} = \sum_{i=1}^{m} I(\mathcal{F}i^{(l)}; \mathcal{R}^{(l)}) + \eta \sum_{i=1}^{m} \sum_{j=i+1}^{m} I(\mathcal{F}_i^{(l)}; \mathcal{F}_j^{(l)}) \tag{5}$$

where m denotes the number of sub-domains in the forget set $\mathcal{F}$, and η is a balancing coefficient that controls the relative importance of inter-domain dependencies. For each layer l, since the activations are high-dimensional and continuous, direct entropy calculation is infeasible [75]. Instead, we utilize Kernel Density Estimation (KDE) to approximate the underlying global data distribution, estimating continuous entropy in activation space as defined in Appendix C [79]. Specifically, we use a multivariate Gaussian kernel, which offers a smooth and flexible density estimation well-suited to high-dimensional data. The estimated probability density function for activations $\mathcal{A}$ is given by:

$$p(a) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{a - a_n}{h}\right) \tag{6}$$

where $a \in \mathbb{R}^d$ represents a single sample from the activations $\mathcal{A}$, including $\mathcal{F}$ and $\mathcal{R}$, with d denoting the feature dimensionality of the activations, N as the number of samples, $K(\cdot)$ represents the kernel function and h as the adaptive bandwidth calculated using Scott's rule [69], defined as $h = \sigma N^{-\frac{1}{d+4}}$, which is particularly suitable for high-dimensional data due to its dimensionality-based adjustment. Here, σ is the standard deviation of the data. This adaptive bandwidth selection effectively balances bias and variance, ensuring robust density estimation for diverse activation distributions [6]. To mitigate the curse of dimensionality, we apply Principal Component Analysis (PCA), which has been widely adopted across various domains in prior work [47, 65, 71] to reduce activation dimensions before performing KDE [2], retaining at least 95% of variance to ensure minimal information loss while significantly lowering computational complexity.

Using the KDE-based entropy estimations, we approximate the overall mutual information $\tilde{I}$ at each layer based on Eq. (5). The optimal layer l^* for unlearning is then determined by minimizing $\tilde{I}$:

$$l^* = \arg\min_{l} \tilde{I}^{(l)} \tag{7}$$

By identifying the layer with the lowest MI, we locate the model region where the *forget* and *retain* datasets are least entangled, minimizing the overlap between the two types of knowledge. Concurrently, this layer exhibits higher entanglement among sub-domains within the *forget* set, enabling efficient updates to shared representations across forgettable sub-domains. This dual property makes the layer an optimal target for unlearning, where parameters with minimal mutual interference are prioritized to remove undesired knowledge while more easily preserving essential and generalizable knowledge for downstream tasks.

4.2 Contrastive Orthogonal Unalignment

To achieve selective knowledge unlearning in LLMs, we first apply MI-guided parameter selection [2] to identify layers with minimal knowledge entanglement, which remains fixed throughout unlearning. We then devise *Contrastive Orthogonal Unalignment* through contrastive mechanisms and gradient projection, employing *alternating strategy* between forget and retain datasets to iteratively refine representations while balancing knowledge removal and retention objectives.

4.2.1 Contrastive Representation Unlearning

The core task of LLM unlearning is to selectively separate knowledge representations to be forgotten from those to be retained. Contrastive learning provides an effective mechanism for this task by

[2] Discussion on MI-guided parameter selection is shown in Appendix. F.2

learning discriminative representations through comparing similar and dissimilar samples. In our context, we leverage contrastive learning to maximize the distance between representations that should be forgotten while maintaining the coherence of retained knowledge.

To facilitate thorough unlearning, we construct Principal Offset Vectors (POVs) that steer model activations away from undesired knowledge by redirecting updated forgettable representations into subspaces intentionally misaligned with the principal directions of frozen counterparts, as identified via SVD, thereby achieving representational decoupling within the model.

Mathematically, given an activation matrix $\mathcal{H} \in \mathbb{R}^{(B \cdot L) \times D}$, where B is the batch size, L the sequence length, and D the hidden dimension, we perform SVD to obtain the dominant principal directions $v_1, \ldots, v_K$ corresponding to the top-K singular values. The POVs $\mathcal{H}^+$ is defined as:

$$\mathcal{H}^+ = \frac{f\left(r \cdot \left(I - w \sum_{i=1}^{K} v_i v_i^\top\right), \epsilon\right)}{\left|f\left(r \cdot \left(I - w \sum_{i=1}^{K} v_i v_i^\top\right), \epsilon\right)\right|} \tag{8}$$

Here, $r \in \mathbb{R}^D$ is a randomly initialized vector, w controls the influence of principal directions, and $I \in \mathbb{R}^{D \times D}$ is the identity matrix. The term ϵ introduces optional perturbations while $f(\cdot)$ is a flexible transformation operator, potentially including non-linear mappings (e.g., tanh), adaptive projections, or adversarially-inspired perturbations, enhancing disentanglement and recovery resistance. This design ensures $\mathcal{H}^+$ is directed away from dominant principal subspaces, combining deterministic guidance and transformations to improve robustness. Unlike generic random vectors, POVs deliberately target dominant features to improve adversarial robustness and unlearning efficacy.

For each input sample, we define three types of representations: the anchor representation $\mathcal{H}_a$ from the updated model for the forget set, the positive representation $\mathcal{H}^+$, given by the POV defined in Eq. (8), and the negative representations $\mathcal{H}^-$ from the frozen model. To ensure consistent scaling, all representations are normalized, and their similarity scores are measured using cosine similarity:

$$S^+ = \sum_{d=1}^{D} \mathcal{H}_a[d] \cdot \mathcal{H}^+[d], \quad S^- = \sum_{z=1}^{\mathcal{Z}} \sum_{d=1}^{D} \mathcal{H}_a[d] \cdot \mathcal{H}_z^-[d] \tag{9}$$

where $\mathcal{Z}$ is the number of negative samples. Building on these similarity scores, we define the forget loss $\mathcal{L}_\mathcal{F}$ using the InfoNCE objective:

$$\mathcal{L}_\mathcal{F} = -\frac{1}{|B|} \sum_{b=1}^{|B|} \log \frac{\exp(S_b^+/\tau)}{\exp(S_b^+/\tau) + \sum_{b=1}^{\mathcal{N}} \exp(S_b^-/\tau)} \tag{10}$$

where τ is a temperature scaling parameter. This loss encourages the updated model's representations to align with the POVs while diverging from the frozen model's representations of undesired knowledge. By leveraging both directional guidance through POVs and contrastive learning, our approach achieves more precise and efficient representation unalignment in activation space.

In addition to unlearning undesired representations, preserving critical knowledge for downstream tasks is essential. We define a retain loss in Eq. (11) to measure alignment between the updated model's activations ($\mathcal{H}^u$) and frozen model's activations ($\mathcal{H}^f$) for the retain set. This retention alignment loss, functioning as a self-supervised variant of contrastive loss, maximizes consistency between updated and frozen activations to ensure effective knowledge preservation during unlearning.

$$\mathcal{L}_\mathcal{R} = 1 - \frac{1}{|B|} \sum_{b=1}^{|B|} \frac{\sum_{d=1}^{D} \mathcal{H}_b^u[d] \cdot \mathcal{H}_b^f[d]}{\sqrt{\sum_{d=1}^{D} \left(\mathcal{H}_b^u[d]\right)^2} \cdot \sqrt{\sum_{d=1}^{D} \left(\mathcal{H}_b^f[d]\right)^2}} \tag{11}$$

This loss ensures alignment between the updated and frozen model activations for the retain set, preserving critical knowledge while complementing the unlearning objective. Combined with the forget loss $\mathcal{L}_\mathcal{F}$, this approach achieves an effective balance between unlearning and retention.

4.2.2 Orthogonalizing Gradient for Conflict Resolution

After computing the forget loss $\mathcal{L}_\mathcal{F}$ and retain loss $\mathcal{L}_\mathcal{R}$, we address optimization direction misalignment between unlearning and retaining by employing a gradient projection mechanism that orthogonalizes conflicting gradients onto subspaces, minimizing interference and promoting balanced optimization. Given the gradients of the forget and retain losses, denoted as $\nabla \mathcal{L}_\mathcal{F}$ and $\nabla \mathcal{L}_\mathcal{R}$, respectively, the conflict can be quantified using the cosine similarity:

$$cos(\nabla\mathcal{L}_{\mathcal{F}}, \nabla\mathcal{L}_{\mathcal{R}}) = \frac{\nabla\mathcal{L}_{\mathcal{F}} \cdot \nabla\mathcal{L}_{\mathcal{R}}}{\|\nabla\mathcal{L}_{\mathcal{F}}\| \cdot \|\nabla\mathcal{L}_{\mathcal{R}}\|} \tag{12}$$

where $cos(\cdot) < 0$ indicates opposing directions, signifying a conflict between the two objectives. To mitigate this conflict, we adjust the gradients by projecting one onto the orthogonal complement of the other. Specifically, if $cos(\cdot) < 0$, we project $\nabla\mathcal{L}_{\mathcal{F}}$ onto the subspace orthogonal to $\nabla\mathcal{L}_{\mathcal{R}}$:

$$\nabla\mathcal{L}_{\mathcal{F}}^{\text{proj}} = \nabla\mathcal{L}_{\mathcal{F}} - \frac{\nabla\mathcal{L}_{\mathcal{F}} \cdot \nabla\mathcal{L}_{\mathcal{R}}}{\|\nabla\mathcal{L}_{\mathcal{R}}\|^2}\nabla\mathcal{L}_{\mathcal{R}} \tag{13}$$

This adjustment ensures that $\nabla\mathcal{L}_{\mathcal{F}}^{\text{proj}}$ is orthogonal to $\nabla\mathcal{L}_{\mathcal{R}}$, eliminating interference from the retain objective during the update for the forget objective. Once the gradients are adjusted, the final update direction of the FALCON is determined by combined gradients:

$$\nabla\mathcal{L}_{FALCON} = \alpha\nabla\mathcal{L}_{\mathcal{F}}^{\text{proj}} + \beta\nabla\mathcal{L}_{\mathcal{R}} \tag{14}$$

where α and β are hyperparameters balancing the contributions of the forget and retain objectives.

This mechanism mitigates gradient conflicts, enabling joint optimization while minimizing interference. By enforcing orthogonality between adjusted gradients, it approximates a Pareto-optimal solution. The model then updates its weights using the conflict-reduced gradient, allowing for more flexible adaptation. To further enhance efficiency and stability, we leverage the second-order optimizer Sophia [52], as suggested in [25, 41], for refined weight updates, ensuring a more effective and stable optimization process for selective knowledge unlearning.

5 Experiments

To validate FALCON's effectiveness, we conduct extensive experiments to answer the following research questions: **RQ1**: Does FALCON with MI guidance, establish a quantifiable measure for principled parameter selection while achieving superior performance in *harmful knowledge unlearning* tasks? (Section 5.1) **RQ2**: Does FALCON maintain strong generalizability across diverse unlearning tasks including *entity unlearning* and *copyrighted content unlearning*? (Section 5.2) **RQ3**: Beyond efficient parameter space reduction through MI guidance, does FALCON's algorithmic design offer competitive *computational efficiency*? (Appendix E.3) **RQ4**: Can FALCON effectively resist *recovery attempts* of unlearned knowledge? (Section 5.3). More complete experiments and ablation study are shown in Appendix E.

5.1 Harmful Knowledge Unlearning

To validate **RQ1**, we use the **WMDP** [50] benchmark for harmful knowledge unlearning assessment, **WikiText** [59] for measuring perplexity, and **MMLU** [26] for evaluating model utility. We test FALCON on three pre-trained LLMs: **Zephyr-7B-Beta** [76], **Yi-6B-Chat** [86], and **Mistral-7B-Instruct-v0.3** [42], comparing against all baselines from [50], with details in Appendix D.

5.1.1 Mutual Information for Parameter Selection

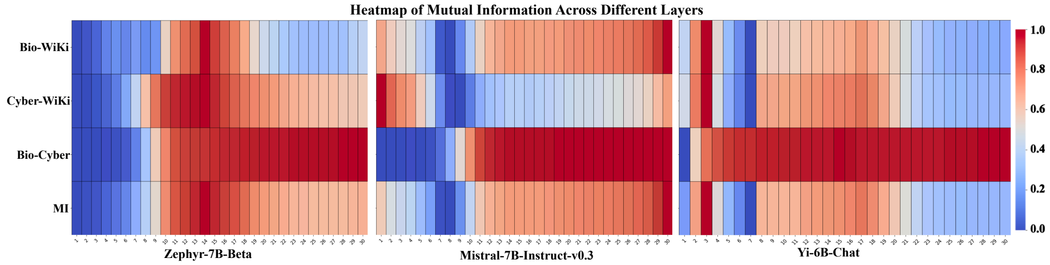

Figure 2: Heatmaps of MI across LLM layers show that lower MI values indicate layers better suited for unlearning, with early layers being more domain-specific and deeper layers more entangled.

Visualization of MI for LLMs Figure 2 presents MI heatmaps illustrating knowledge entanglement between forget sets (WMDP-Bio, WMDP-Cyber) and the retain set (WikiText-2-raw-v1) across LLM layers. This metric provides an interpretable measure for identifying layers with minimal entanglement for targeted unlearning. All models show lower MI values in earlier layers, indicating more domain-specific and disentangled representations, which aligns with both intuition and experimental observations [50]. Yi-6B-Chat demonstrates particularly complex entanglement patterns between domains, presenting a greater difficulty for unlearning multi-domain knowledge and making it an ideal candidate for our effectiveness analysis experiments in Section 5.1.2. Beyond identifying optimal intervention parameters, MI-guided selection improves efficiency by narrowing the parameter search space compared to exhaustive methods like grid search, scaling effectively with model complexity.

Gradient Conflicts Analysis We empirically validate the underlying principle of MI guidance by analyzing gradient conflicts between forget and retain objectives across layers. As shown in Figure 3, layers with low MI values exhibit significantly reduced conflicts, with cosine similarities near zero, indicating minimal interference between objectives. Conversely, high-MI layers show pronounced, fluctuating conflicts, highlighting the issues of entangled representations. These results confirm that mutual information is a reliable auxiliary signal for guiding parameter selection, as low-MI parameters reduce interference, support stable updates, and help mitigate conflicts between unlearning and retention goals.

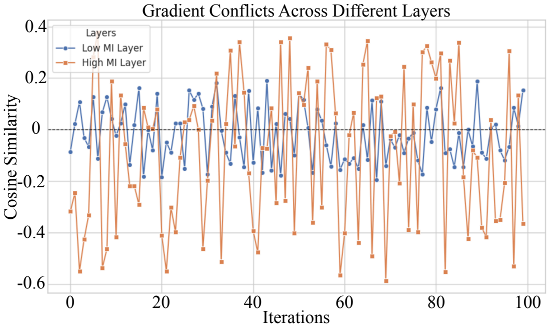

Figure 3: Gradient conflicts across layers with minimum (blue) and maximum (orange) MI values computed during parameter selection in Mistral-7B.

5.1.2 Unlearning Effectiveness and Utility Analysis

Table 1: Unlearning effectiveness and utility across models and methods. Metrics with (↑) indicate preferable increases; (↓) indicate preferable decreases.

Method	**WMDP (↓)**		**MMLU (↑)**	**PPL (↓)**
	Bio	**Cyber**		
Zephyr-7B	63.7	43.8	58.1	1.5
+ LLMU	36.3	40.5	50.3	4.8
+ SCRUB	38.7	35.4	50.0	16.5
+ SSD	53.1	43.2	52.8	1.6
+ RMU	34.5	28.9	57.4	1.5
+ FALCON	**26.7**	**25.3**	**57.4**	**1.5**
Yi-6B-Chat	65.4	42.6	61.8	1.5
+ LLMU	56.2	39.9	57.5	5.4
+ SCRUB	38.7	35.5	50.0	16.4
+ SSD	55.1	43.7	53.8	1.6
+ RMU	50.8	33.5	59.6	1.6
+ FALCON	**27.7**	**25.3**	**60.3**	**1.5**

We evaluate FALCON against all baseline methods across three LLM architectures shown in Table 1 and Appendix E.1, with our evaluation focusing on three key metrics: WMDP scores for measuring unlearning effectiveness, MMLU scores for assessing general knowledge retention, and perplexity (PPL) for model stability. Our primary objective is to *minimize WMDP scores while maintaining MMLU and PPL values close to the base model's performance (MMLU and PPL)*, as this indicates successful knowledge removal without compromising general capabilities. To ensure quantifiable comparison, we prioritize maintaining general model utility and report each method's best unlearning performance under this setting. Results demonstrate FALCON's superior performance compared to baselines that struggle with effectiveness-utility balance and show increased uncertainty in their perplexity. On Zephyr-7B, FALCON achieve lower forgetting scores while preserving general capabilities. This advantage is more clear on Yi-6B-Chat with its complex knowledge entanglement: RMU show significant biological domain degradation when constrained to maintain MMLU above 60%, while FALCON maintain consistent effectiveness with superior general performance. These findings validate our fine-grained representation-guided mechanisms for targeted unlearning with preserved utility, even in scenarios with complex knowledge entanglement.

5.2 Cross-Domain Generalizability Assessment

To address **RQ2**, we conduct additional experiments on copyrighted content and entity unlearning using the **MUSE** [72] and **TOFU** [56] benchmarks with additional baselines [16]. For **RQ3**, we compare computational efficiency across methods in Appendix E.3). All aforementioned experiments utilize *first-order optimizers for fair comparison*, with complete implementation details in Appendix D

5.2.1 Copyrighted Content Unlearning

For copyrighted content unlearning, we utilize the MUSE benchmark and Llama-2-7b-hf to assess FALCON's effectiveness in removing protected news articles while preserving general capabilities. As shown in Table 2, FALCON achieved the lowest forget metrics scores (0.02 and 0.03) while maintaining competitive retention (0.54). Unlike baselines, FALCON consistently balanced copyright removal with knowledge preservation, demonstrating broader applicability beyond harmful content removal.

Table 2: Evaluation on MUSE News over 10 epochs.

Method	**forget_knowmem_ROUGE↓**	**forget_verbmem_ROUGE↓**	**retain_knowmem_ROUGE↑**
Finetuned	0.64	0.58	0.55
Retain	0.33	0.21	0.56
GradAscent	0.00	0.00	0.00
GradDiff	0.41	8.92e-3	0.37
NPO	0.56	0.35	0.51
SimNPO	0.54	0.36	0.51
RMU	0.48	0.05	0.51
FALCON	**0.02**	**0.03**	**0.54**

5.2.2 Entity Unlearning

We evaluate FALCON's ability to remove knowledge about fictitious entities using TOFU with varying forget data sizes (1/5/10%). Our method maintain strong forget quality (FQ↑) and model utility (MU↑) across different splits on Llama-3.2-1B-Instruct. Even with only 10 unlearning epochs, FALCON consistently outperform baselines in balancing knowledge removal with preserved utility. Notably, while other methods like GradAscent suffers significant utility degradation with larger forget sets, FALCON remains effective, demonstrating our method's generalizability to entity unlearning tasks.

Table 3: TOFU evaluation across varying sizes over 10 epochs.

Method	Forget01		Forget05		Forget10	
	FQ	MU	FQ	MU	FQ	MU
Finetuned	0.01	0.60	2.96e-13	0.60	8.08e-22	0.6
Retain	1.0	0.60	1.0	0.60	1.0	0.59
GradAscent	0.27	0.33	1.94e-119	0	1.06e-239	0
GradDiff	0.77	0.43	2.04e-110	0.22	1.06e-239	0.49
IdkDPO	0.01	0.51	4.02e-06	0.04	4.26e-10	0.08
NPO	0.92	0.56	0.32	0.42	0.02	0.46
RMU	0.16	0.55	1.46e-7	0.57	1.4e-20	0.59
FALCON	**0.99**	**0.55**	**0.92**	**0.59**	**0.52**	**0.60**

5.3 Resistance Against Knowledge Recovery Attempts

We conduct experiments on Yi-6B-Chat to evaluate FALCON's resistance against knowledge recovery attempts [55] for **RQ4**. Logit Lens [61], which projects intermediate activations onto the model's vocabulary space, serves as a powerful technique for probing the model's internal knowledge representations and potential recovery of unlearned information. As shown in Figure 4, the logit lens analysis across different architectural components such as MLP and attention layers demonstrates that the unlearned knowledge remains consistently inaccessible, with performance staying close to the unlearned baseline and far below the original model's performance.

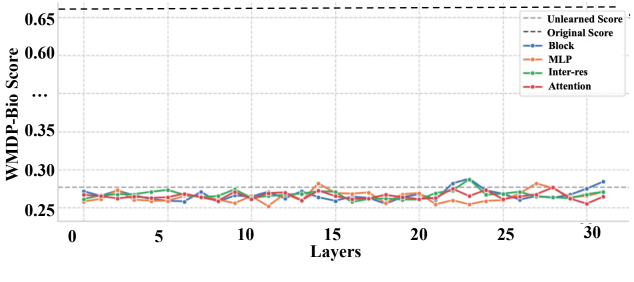

Figure 4: Logit lens probing results on different components.

Additionally, as shown in Table 4, FALCON exhibits strong resilience against enhanced GCG in QA setting, an advanced prefix-optimization based jailbreaking attack that compromises other baselines such as RMU [73]. Even with increasing attack iterations, the recovered WMDP scores remain close to the unlearned baseline, demonstrating robust unlearning through fundamental changes to the model's internal representations rather than superficial knowledge masking. Further evaluation using conversational templates for jailbreaking attacks (detailed in Appendix E.6) further validates our method's robustness against knowledge recovery attempts. These results across both probing techniques validate FALCON's effectiveness in creating a more permanent and recovery-resistant form of knowledge removal.

Table 4: Knowledge recovery results via enhanced GCG attack.

Dataset	Original Score	Unlearning Score	Recovery Score via Enhanced GCG			
			GCG-500	GCG-1000	GCG-1500	GCG-2000
WMDP-Bio	65.4	27.7	27.6	28.4	27.9	28.9
WMDP-Cyber	42.6	25.3	26.3	26.4	25.8	24.7

6 Practical Implications of LLM Unlearning for Responsible AI

The problem setting addressed by FALCON[3] stems from the growing challenge of directly employing LLMs or deploying them as autonomous agents in safety-critical environments [32, 34]. As these models become increasingly embedded in diverse real-world applications, selectively removing undesired or harmful knowledge after deployment remains difficult [54]. Unlike conventional machine learning models where unwanted data can simply be excluded in future training cycles, LLMs encode information across billions of parameters, making precise removal extremely challenging. This limitation creates a critical gap between learning capabilities and responsible deployment. The issue is further amplified by regulatory demands such as the GDPR's "right to be forgotten" [67], and by empirical evidence that even state-of-the-art LLMs and their agentic variants can inadvertently reproduce sensitive or hazardous content when prompted, raising urgent concerns about information safety and controllability.

FALCON provides a fine-grained unlearning mechanism that identifies harmful knowledge and decouples it from beneficial reasoning. This targeted process enables models to forget unsafe information while retaining legitimate competence, supporting the emerging need for responsible LLM deployment [58, 31]. As LLMs operate in dynamic, real-world contexts, the capacity for precise and interpretable knowledge modification becomes essential for responsible AI. We advocate viewing unlearning not as an academic objective but as core practical infrastructure for transparent, compliant, and responsible AI systems.

7 Conclusion

This paper presents FALCON, a fine-grained representation-guided framework for LLM unlearning. Leveraging mutual information guidance and contrastive orthogonal unalignment, it enables precise and efficient unlearning through principal component-based representation separation and gradient conflict resolution. Extensive experiments demonstrate its superior performance in effectively removing undesired knowledge while preserving essential information across diverse tasks, along with resistance against knowledge recovery and efficient optimization guidance. However, this work is currently limited to text-based LLM unlearning, with experiments conducted on relatively smaller models due to computational constraints. Future directions include extending unlearning to multimodal LLMs and refining strategies to disentangle intertwined knowledge in deeper architectures.

Acknowledgment

This work is partially funded by the European Union (under grant agreement ID 101212818). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work is partially supported by Innovate UK through AI-PASSPORT under Grant 10126404. Yi's contribution is partially supported through the Royal Society international exchanges programme and in part by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from RAi UK.

[3] Additional discussions are provided in Appendix F.

References

[1] 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.

[2] Naomi Altman and Martin Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15(6):399–400, 2018.

[3] Shuang Ao, Yi Dong, Jinwei Hu, and Sarvapali Ramchurn. Safe pruning lora: Robust distance-guided pruning for safety alignment in adaptation of llms. *arXiv preprint arXiv:2506.18931*, 2025.

[4] Giuseppe Attanasio, Debora Nozza, Dirk Hovy, Elena Baralis, et al. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119. Association for Computational Linguistics, 2022.

[5] Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. Open problems in machine unlearning for ai safety, 2025.

[6] Elsidieg I Belhaj. A modified rule-of-thumb method for kernel density estimation. 2024.

[7] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European conference on computer vision*, pages 440–457. Springer, 2022.

[8] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11186–11194, 2024.

[9] Cheng Chen, Ji Zhang, Jingkuan Song, and Lianli Gao. Class gradient projection for continual learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5575–5583, 2022.

[10] Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. Learning to maximize mutual information for chain-of-thought distillation. *arXiv preprint arXiv:2403.03348*, 2024.

[11] Wenying Deng, Beau Coker, Rajarshi Mukherjee, Jeremiah Liu, and Brent Coull. Towards a unified framework for uncertainty-aware nonlinear variable selection with theoretical guarantees. *Advances in Neural Information Processing Systems*, 35:27636–27651, 2022.

[12] Ann-Kathrin Dombrowski and Guillaume Corlouer. An information-theoretic study of lying in LLMs. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

[13] Yi DONG, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Position: Building guardrails for large language models requires systematic design. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*, 2024.

[15] Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. *arXiv preprint arXiv:2402.10052*, 2024.

[16] Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, and Pratyush Maini. OpenUnlearning: A unified framework for llm unlearning benchmarks. `https://github.com/locuslab/open-unlearning`, 2025. Accessed: February 27, 2025.

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

[19] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[20] 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.

[21] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[22] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12043–12051, 2024.

[23] Marylou Gabrié, Andre Manoel, Clément Luneau, Nicolas Macris, Florent Krzakala, Lenka Zdeborová, et al. Entropy and mutual information in models of deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[24] Piotr Garbaczewski. Differential entropy and dynamics of uncertainty. *Journal of Statistical Physics*, 123:315–355, 2006.

[25] Kang Gu, Md Rafi Ur Rashid, Najrin Sultana, and Shagufta Mehnaz. Second-order information matters: Revisiting machine unlearning for large language models. *arXiv preprint arXiv:2403.10557*, 2024.

[26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[27] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[28] Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*, 2024.

[29] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[30] Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. A comprehensive survey on contrastive learning. *Neurocomputing*, page 128645, 2024.

[31] Jinwei Hu, Yi Dong, Shuang Ao, Zhuoyun Li, Boxuan Wang, Lokesh Singh, Guangliang Cheng, Sarvapali D Ramchurn, and Xiaowei Huang. Position: Towards a responsible llm-empowered multi-agent systems. *arXiv preprint arXiv:2502.01714*, 2025.

[32] Jinwei HU, Yi DONG, Zhengtao DING, and Xiaowei HUANG. Enhancing robustness of llm-driven multi-agent systems through randomized smoothing. *Chinese Journal of Aeronautics*, page 103779, 2025.

[33] Jinwei Hu, Yi Dong, and Xiaowei Huang. Trust-oriented adaptive guardrails for large language models. *arXiv preprint arXiv:2408.08959*, 2024.

[34] Jinwei Hu, Yi Dong, Youcheng Sun, and Xiaowei Huang. Tapas are free! training-free adaptation of programmatic agents via llm-guided program synthesis in dynamic environments. *arXiv preprint arXiv:2508.11425*, 2025.

[35] James Y Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *arXiv preprint arXiv:2404.11045*, 2024.

[36] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

[37] Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, July 2023.

[38] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023.

[39] Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611, 2024.

[40] 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. *arXiv preprint arXiv:2410.17509*, 2024.

[41] Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*, 2024.

[42] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[43] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*, 2024.

[44] Hongpeng Jin, Wenqi Wei, Xuyu Wang, Wenbin Zhang, and Yanzhao Wu. Rethinking learning rate tuning in the era of large language models. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence*, pages 112–121, 2023.

[45] Jonathan Kahana and Yedid Hoshen. A contrastive objective for learning disentangled representations. In *European Conference on Computer Vision*, pages 579–595. Springer, 2022.

[46] Junyaup Kim and Simon S Woo. Efficient two-stage model retraining for machine unlearning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4361–4369, 2022.

[47] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270. PMLR, 2023.

[48] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36, 2024.

[49] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202. Association for Computational Linguistics, November 2020.

[50] 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 *Forty-first International Conference on Machine Learning*, 2024.

[51] Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266, 2024.

[52] Hong Liu, Zhiyuan Li, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, and Zhan Qin. Muter: Machine unlearning on adversarially trained models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4892–4902, 2023.

[54] 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*, pages 1–14, 2025.

[55] Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. In *Transactions on Machine Learning Research*, 2024.

[56] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for llms. In *First Conference on Language Modeling*, 2024.

[57] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[58] Kim Martineau. Why we're teaching llms to forget things. `https://research.ibm.com/blog/llm-unlearning`, October 2024.

[59] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[61] Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

[62] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning*, 2024.

[63] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

[64] Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[65] Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay Cohen. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

[66] Youyang Qu, Ming Ding, Nan Sun, Kanchana Thilakarathna, Tianqing Zhu, and Dusit Niyato. The frontier of data erasure: Machine unlearning for large language models. *arXiv preprint arXiv:2403.15779*, 2024.

[67] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

[68] Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024.

[69] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[70] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2024.

[71] Shun Shao, Yftah Ziser, and Shay B. Cohen. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[72] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. 2024.

[73] T Ben Thompson and Michael Sklar. Flrt: Fluent student-teacher redteaming. *arXiv preprint arXiv:2407.17447*, 2024.

[74] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

[75] Dor Tsur, Ziv Goldfeld, and Kristjan Greenewald. Max-sliced mutual information. *Advances in Neural Information Processing Systems*, 36, 2024.

[76] 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.

[77] Aleksandra Urman and Mykola Makhortykh. The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat. 2023.

[78] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[79] Janett Walters-Williams and Yan Li. Estimation of mutual information: A survey. In *Rough Sets and Knowledge Technology: 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14-16, 2009. Proceedings 4*, pages 389–396. Springer, 2009.

[80] Jiaan Wang, Jianfeng Qu, Kexin Wang, Zhixu Li, Wen Hua, Ximing Li, and An Liu. Improving the robustness of knowledge-grounded dialogue via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19135–19143, 2024.

[81] Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[82] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8496–8504, 2025.

[83] Tzu-Hsuan Yang and Cheng-Te Li. When contrastive learning meets graph unlearning: Graph contrastive unlearning for link prediction. In *2023 IEEE International Conference on Big Data*, pages 6025–6032. IEEE, 2023.

[84] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Socially Responsible Language Modelling Research*, 2023.

[85] Ziyi Yin, Muchao Ye, Tianrong Zhang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vqattack: Transferable adversarial attacks on visual question answering via pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6755–6763, 2024.

[86] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[87] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, pages 1–10, 2024.

[88] Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023.

[89] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

[90] Qiuchen Zhang, Carl Yang, Jian Lou, Li Xiong, et al. Contrastive unlearning: A contrastive approach to machine unlearning. *arXiv preprint arXiv:2401.10458*, 2024.

[91] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

[92] Zuyu Zhang, Yan Li, and Byung-Seok Shin. Embracing domain gradient conflicts: Domain generalization using domain gradient equilibrium. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5594–5603, 2024.

[93] Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[94] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. *arXiv preprint arXiv:2402.11889*, 2024.

[95] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the main contributions of FALCON, including its practicality and effectiveness.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper discusses current limitations in conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work present many experimental evidences and it is primarily empirical in nature.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of datasets, model configurations, and evaluation metrics to facilitate reproducibility.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized GitHub repository will be released upon acceptance, containing implementation and experiment code

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training setups, hyperparameters, and baseline configurations are thoroughly documented in the main paper and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Due to the intrinsic non-deterministic behavior of LLM, traditional significance testing is not directly applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the use of specific models, sizes and prove the unlearning efficiency in appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with NeurIPS ethical standards and focuses on enhancing model safety through principled unlearning.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both the benefits of preventing misuse through knowledge removal.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work does not involve the release of high-risk models or datasets that require specific safety measures.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, models, and code used from prior work are properly cited and gain consent.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The code will be released upon acceptance due to double-blind review constraints.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve human subjects or crowdsourced data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.