Shengming Zhang

michaelzhang@ibms.pumc.edu.cn Chinese Academy of Medical Sciences & Peking Union Medical College Beijing, China

Le Zhang\* zhangle09@baidu.com Baidu Research Beijing, China

Jingbo Zhou zhoujingbo@baidu.com Baidu Research Beijing, China

Zhi Zheng zhengzhi97@mail.ustc.edu.cn University of Science and Technology of China Hefei, China

## Abstract

We focus on unlearning unwanted knowledge in autoregressive large language models (LLMs) through pruning. Our goal is to selectively remove undesirable information (e.g., harmful responses, privacy-sensitive data) while ensuring the preservation of desirable knowledge (e.g., positive responses and objective facts). Previous approaches use gradient ascent (GA) over undesired knowledge to inversely optimize LLMs, which compromises the model's performance on desired knowledge. To address this limitation, we introduce a novel two-stage approach, named LLM-Eraser, for selectively identifying and editing parameters specifically associated with undesirable knowledge. LLM-Eraser operates in two stages: localization and unlearning. During the localization stage, we utilize neuron scores and trainable soft masks to identify parameters crucial to the undesired knowledge. In the unlearning stage, we prune these identified parameters and apply a selective post-training process to enhance the model's selectiveness. Our experiments, conducted across five task datasets, demonstrate that LLM-Eraser effectively unlearns undesirable knowledge-evidenced by the model's nearrandom performance on multiple-choice questions related to the erased knowledge-while maintaining high proficiency in desirable knowledge, with an average performance deficit of only 2.5%.

## **CCS** Concepts

• Computing methodologies → Machine learning; Natural language generation; • Security and privacy; • Theory of computation;

\*Le Zhang and Hui Xiong are both Corresponding Authors.

KDD '25, August 3-7, 2025, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-1245-6/25/08

https://doi.org/10.1145/3690624.3709312





Figure 1: Illustration of LLM modification techniques.

## Keywords

Machine Unlearning, Large Language Models, Selective Pruning, Model Finetuning

#### **ACM Reference Format:**

Shengming Zhang, Le Zhang, Jingbo Zhou, Zhi Zheng, and Hui Xiong. 2025. LLM-Eraser: Optimizing Large Language Model Unlearning through Selective Pruning. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3-7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 12 pages. https://doi.org/ 10.1145/3690624.3709312

#### 1 Introduction

Large Language Models (LLMs) have demonstrated impressive advancements in language comprehension and synthesis [7, 10, 16, 43], leading to emergent abilities such as multi-modality [22, 48, 56] and chain-of-thought reasoning [46], among others [20, 45, 49, 60]. However, the exceptional learning capabilities of LLMs also raise ethical and safety concerns, leading to potential misuse, such as academic dishonesty [13], fabricating restaurant reviews [1], and facilitating malicious activities like cyberattacks or bioweapon development [29]. These concerns highlight the necessity of adaptively editing LLMs to mitigate misuse while preserving their ability to respond to normal prompts.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.



# Figure 2: Comparison of LLM-Eraser with existing approach and original LLM on desired and undesired knowledge. Existing gradient ascent-based approach harms performance over desired knowledge.

Mitigating misuse of LLMs can be optimistically achieved by modifying the models so that they avoid generating responses associated with undesirable knowledge. Figure 1 illustrates several common LLM modification techniques. General-task pruning (i) is primarily focused on model compression and does not address the goal of misuse mitigation. Techniques like Knowledge Distillation (ii) and LLMs with adapters (iii) can help mitigate misuse, but they either require an extensive retraining process [37, 40] for distillation or rely on removable adapters, leaving the fixed LLMs still vulnerable to misuse.

Alternatively, directly erasing memories associated with undesired knowledge, known as Machine Unlearning (MU) [5, 8, 36], offers a promising solution. Unlearning in LLMs has been explored to mitigate issues like toxicity [25], copyright and privacy violations [19], fairness [53], and sensitive knowledge [3, 23]. However, existing approaches often struggle to balance eliminating undesired knowledge and maintaining performance on normal inputs [23, 24, 52].

Our objective is to effectively unlearn undesired memories in LLMs while safeguarding essential, desired knowledge with minimal performance compromise. We conclude to address two significant challenges in our paper:

- Selective identification and erasure of undesired knowledge: Existing unlearning techniques focus on negating undesired knowledge based on specific data samples. Balancing the erasure of undesired knowledge and the preservation of desired knowledge requires precise localization of model parameters closely associated with the undesired knowledge, which is inherently challenging [24].
- Knowledge Entanglement in LLMs: Knowledge in LLMs is intricately interlaced, with significant portions stemming from clusters of crucial neurons. Unlearning these neurons could lead to catastrophic performance impacts due to the interconnected nature of the knowledge [23]. Hence, even after locating memories tied to undesired knowledge, modifying these memories adversely affect desired knowledge.

To address the complexities of unlearning in LLMs, we propose a novel methodology called LLM-Eraser, which could fast identify knowledge associated with undesired knowledge and unlearns it with selective pruning (Figure 1 (iv)). LLM-Eraser employs a twostep process: localization and unlearning. In the localization stage, we explore both score-based and mask-based techniques to identify undesired memories [26, 53]. Score-based localization uses a second-order Taylor expansion to approximate changes in the loss function relative to model parameters, while mask-based localization employs a trainable soft mask with *l*-1 regularization to enforce sparsity. In the unlearning stage, we prune a portion of the model parameters identified during localization and apply a selective tuning process to enhance performance. The selective tuning process erases undesired knowledge by enforcing a random mismatch and preserves desired knowledge by minimizing divergence from the original LLM. Additionally, a contrastive objective is introduced to disentangle desired and undesired knowledge. Figure 2 shows comparison of our proposed LLM-Eraser with existing gradient ascent-based approach and original LLM, showcasing the effectiveness of LLM-Eraser in balancing the negation and preservation.

Our experiments, conducted across five general-task benchmark datasets and specific language generation corpus datasets, demonstrate the effectiveness of LLM-Eraser. The results show that LLM-Eraser successfully eliminates undesired knowledge while largely retaining comparable performance for desired knowledge.

## 2 Related Work

### 2.1 Large Language Model Unlearning

LLMs are derived from transformer architectures, which have already proven versatile across various applications [54, 57-59], further extending their impact through the capabilities of LLMs. We focus on model-based machine unlearning approaches over large language models. These approaches involve the modification of LLMs' weights and/or structures for the unlearning purpose [19, 25, 38, 52]. One of the commonly used model-based machine unlearning techniques is Gradient Ascent (GA)-based approaches, which optimize LLMs to the inversed gradient directions w.r.t. undesired knowledge [19, 52]. GA-based approaches can be sensitive to the training procedures, and direct gradient ascent over undesired knowledge leads to a noticeable performance drop on desired knowledge [15]. To stabilize the GA-based machine unlearning process, one of the variants transforms it into a gradient descent problem, which aims at maximizing the likelihood of predictions on undesired knowledge with relabeled (e.g., randomly relabeled) data [53]. However, these variants focus primarily on erasing undesired knowledge and overlook the balance to preservation.

## 2.2 Comparison with Model Editing Approaches

Model editing is a closely related topic to model unlearning, which focuses on editing factual associations incorporated within LLMs [14, 30, 31] or to mitigate fictitious knowledge [17, 28, 42]. A model editing approach can be mitigated for the knowledge erasure task by substituting the factual association labels, and it is also crucial for model editing approaches to edit factual associations within a specific scope [12, 33]. Model editing can be approached by localizing a subset of model units that are closely related to the scope to be edited [9, 47], which is instructive for knowledge erasure as well. However, model editing differs from model unlearning: The primary focus of model unlearning is to erase memories, while model editing modifies or integrates with new memories, which does not focus on elimination. The mitigation of model editing approaches for the unlearning task forces LLMs to substitute the answer to undesired knowledge with pre-defined sentences, e.g. spaces or fixed refuse-to-answer context [27, 33, 51]. However, an adversarial can easily obtain undesired knowledge through logit scores since the relative tendency has not been completely erased, only substituted.

## 3 Problem Statement

#### **Table 1: Mathematical Notations**

Symbol	Description
$\mathcal{F}_{\Theta}(\cdot)$	LLM as a parameterized projection function
X	Prompt input
У	Generated outputs
Θ	Parameter set of LLM
$D(\cdot, \cdot)$	Pre-defined distance metrics, e.g. cross-entropy loss
Χ	$X = \{X_1, X_2,, X_n\}$ Set of prompt input genres
$\mathcal{F}^*_{\Theta}(\cdot)$	Optimal LLM with perfect outputs
$\mathcal{F}_{\Theta}'$	Selective LLM
REJECT	Reject operations for undesired knowledge
θ	$\theta \in \Theta$ a subset of parameters of $\Theta$
$\mathcal{L}_{ heta}$	Loss function value w.r.t. $\theta$
$I_{\Theta}$	Importance score of parameters
Н	Hessian matrix of Importance score
m	Pruning mask of parameters
$\mathcal{D}_d$	Dataset of desired knowledge
$\mathcal{D}_u$	Dataset of undesired knowledge
$\alpha, \beta, \alpha', \beta'$	Scaling factors
$\lambda_{train}, \lambda_{div}$	Scaling factors
z	Output logits of LLM
$\hat{z}$	$\hat{z} = \{z_1, z_2,, z_N\}$ selected anchors to contrast
Τ, τ	Temperature hyperparameter

We first give a formal problem statement of the undesired knowledge erasure task. For the purpose of clarity and comprehensibility, we encapsulate the mathematical notations used throughout this paper in Table 1. A Large Language Model can be regarded as a parameterized projection function  $\mathcal{F}_{\Theta}(\cdot)$  that projects prompt input Xinto answer space  $\mathcal{Y}$ . Parameter set  $\Theta$  is optimized to minimize the distance between the projected output and the answer label, formally:

$$\min_{\Theta} D(\mathcal{F}_{\Theta}(\mathcal{X}), \mathcal{Y}), \tag{1}$$

where  $D(\cdot, \cdot)$  can be arbitrary pre-defined distance metrics.

KDD '25, August 3-7, 2025, Toronto, ON, Canada

Given a set of prompt input genres  $X = \{X_1, X_2, ..., X_n\}$ , where  $X_i$  indicates any prompt input depending on a certain **genre of knowledge**, e.g. coding or chemistry. Without lost of generality, let the first *m* genres of prompt inputs in X depend on **desired knowledge**, e.g. objective facts, and the remaining genres of prompt inputs depend on **undesired knowledge**, e.g. knowledge to make biochemical weapons. Formally, given all genres of knowledge  $X = \{X_1, X_2, ..., X_m, ..., X_n\}, \forall i \leq m, X_i$  depends on desired knowledge, otherwise depending on undesired knowledge. An optimal general-task LLM  $\mathcal{F}_{\Theta}^{=}(\cdot)$  could precisely answer prompt inputs of any genre regardless depending on desired or undesired knowledge:

$$D(\mathcal{F}^*_{\Theta}(\mathcal{X}_i), \mathcal{Y}_i) \to 0, \ \forall \mathcal{X}_i \in \mathcal{X}.$$
(2)

For the task of selective undesired knowledge erasure, we aim to build a selective model  $\mathcal{F}'_{\Theta}(\cdot)$  based on the optimal generaltask LLM, such that the selective model could optimally answer any genres of prompt inputs depending on desired knowledge, yet **rejects** answering any prompt input coming from genres depending on undesired knowledge. Formally:

$$\mathcal{F}_{\Theta}'(X_i) \to \begin{cases} \mathcal{Y}_i, & i \le m \\ \mathbf{REJECT}, & i > m. \end{cases}$$
(3)

Here **REJECT** can be any operations irrelevant to ground-truth label  $\mathcal{Y}_i$ , e.g., random text, fixed content, or empty.

## 4 Methodology

In this section, we unpack the specifics of LLM-Eraser designed for the selective knowledge erasure task. Generally, LLM-Eraser operates iteratively alternating between two central stages: the localization stage and the unlearning stage. Within the localization stage, LLM-Eraser singles out parameters associated with high relative importance in relation to undesired knowledge, while maintaining moderate importance for the desired knowledge. Transitioning to the unlearning stage, a particular portion of these identified parameters is pruned, followed by a selective tuning process to counter any prospective performance decline caused by the preceding pruning. The selective tuning process enforces the negation of undesired knowledge using a random mismatch objective, and enforces utilization of desired knowledge by minimizing the divergence between the original LLM. In order to decompose the correlation between desired and undesired knowledge, we put forth a contrastive objective to enforce disentanglement between desired and undesired knowledge. Figure 3 shows the architecture of LLM-Eraser.

#### 4.1 Mathematical Modeling

Building upon the problem statement in Section 3, we propose the mathematical modeling for selective knowledge erasure [23]:

$$\underset{\Theta}{\min} \underbrace{ \underbrace{\mathbf{E}_{(x,y)\in\mathcal{D}_d}[l(y_d|x;\Theta)]}_{\text{preserve}} + \underbrace{\lambda \mathbf{E}_{(x,y)\in\mathcal{D}_u}[l(y_u|x;\Theta)]}_{\text{eliminate}}, \quad (4)$$

where  $\mathcal{D}_d$ ,  $y_d$  and  $\mathcal{D}_u$ ,  $y_u$  correspond to data and label of desired and undesired knowledge. The first term enforces preservation of LLM's ability for desired knowledge, and the second term enforces elimination of undesired knowledge. KDD '25, August 3-7, 2025, Toronto, ON, Canada

Shengming Zhang, Le Zhang, Jingbo Zhou, Zhi Zheng, and Hui Xiong



Figure 3: Architecture of proposed LLM-Eraser framework. Localization of activated neurons and weights for desired and undesired knowledge are shown in a) and b); Neurons exclusively activated by undesired knowledge are pruned as shown in c). The selective tuning process has three loss terms, as shown in d).

## 4.2 Localization Stage

We propose two selective localization variants - score-based localization and mask-based localization - specifically tailored for the undesired knowledge selective erasure task.

4.2.1 Score-based Localization. Score-based localization derives the importance score to indicate localization. Given an LLM loss function  $\mathcal{L}_{\Theta}$  w.r.t. model parameters  $\Theta$ , the importance score of parameter, i.e.  $I_{\Theta}$  can be determined as the change of loss function value if setting the parameter as zero. Formally:

$$\Theta = |\mathcal{L}_{\Theta} - \mathcal{L}_{\Theta=0}|. \tag{5}$$

The change of loss function values w.r.t. model parameters can be approximated by Taylor expansion [34]. Early research approximates it using first-order expansion, i.e. the gradient of parameters [50]. However, under a well-trained scheme, the gradient of model parameters can be mostly zero, thus we adopt the secondorder expansion in approximating the importance score. Formally:

$$I_{\Theta} \approx \left| \frac{\partial \mathcal{L}_{\Theta}}{\partial \Theta} \Theta + \frac{1}{2} \Theta^T H \Theta + O(||\Theta||^3) \right|, \tag{6}$$

where *H* is the Hessian matrix,  $O(||\Theta||^3)$  is a third-order quantity that can be ignored. We define the second-order selective localization importance score as:

$$I_{\Theta} \approx \mathbb{E}_{(x,y)\in\mathcal{D}_{u}} \left| \frac{\partial \mathcal{L}(x;\Theta,y)}{\partial \Theta} \Theta + \frac{1}{2} \Theta^{T} H \Theta \right| - \alpha \mathbb{E}_{(x,y)\in\mathcal{D}_{d}} \left| \frac{\partial \mathcal{L}(x;\Theta,y)}{\partial \Theta} \Theta + \frac{1}{2} \Theta^{T} H \Theta \right| ,$$
(7)

where  $\alpha$  is a scaling factor. The selective importance score assigns a much higher score to neurons important only to undesired knowledge genres, while keeping moderate score for neurons important to both desired and undesired knowledge genres.

4.2.2 *Mask-based Localization*. Mask-based localization optimizes masks for model parameters in identification of crutial parameters [9, 39, 47]. Formally, the objective for optimizing the mask **m** can be written as:

$$\min_{\mathbf{m}\in\{0,1\}^n} \mathbf{E}_{(x,y)\in\mathcal{D}}\mathcal{L}(x;\mathbf{m}\odot\Theta,y),\tag{8}$$

where  $\odot$  is the element-wise product between mask **m** and parameter  $\Theta$ . By minimizing the impact of masking out a group of model parameters, mask-based localization could identify parameters (not being masked) crucial for maintaining consistency of model output. Directly optimizing the mask within binary range makes the optimization process discontinuous. Following [9, 50], we relax the mask into real numbers, i.e.  $\mathbf{m} \in [0, 1]^n$ , and add the *l*-1 regularization to enforce sparsity. Formally:

$$\min_{\mathbf{m}\in[0,1]^n} \mathbf{E}_{(x,y)\in\mathcal{D}}\mathcal{L}(x;\mathbf{m}\odot\Theta,y) + \beta ||\mathbf{m}||_1,$$
(9)

where (x, y) is sampled data from dataset  $\mathcal{D}$ ,  $\beta$  is a hyperparameter to control the regularization strength. In order to selectively identify undesired knowledge, we propose a selective mask-based localization scheme. Formally:

$$\min_{\mathbf{m}\in[0,1]^n} \quad \mathbb{E}_{(x,y)\in\mathcal{D}_d}\mathcal{L}(x;\mathbf{m}\odot\Theta,y) - \alpha\mathbb{E}_{(x,y)\in\mathcal{D}_u}\mathcal{L}(x;\mathbf{m}\odot\Theta,y) + (10)$$
$$\beta||\mathbf{m}||_1.$$

The mask-based selective localization objective locates (not being masked) parameters positively related to desired knowledge, while negatively related to undesired one, and ensures sparsity.

## 4.3 Unlearning Stage

The localization stage selectively identifies crucial parameters exclusively related to undesired knowledge. In the unlearning stage, we discuss how to eliminate undesired knowledge while preserving desired knowledge with the guidance of localization. Once we have obtained the importance score or learned masks of parameters during the localization stage, we prune out certain percentage of parameters that are most relevant to undesired knowledge. Specifically, for score-based localization, we prune parameters w.r.t. Equation 7 from high to low, and for mask-based localization, we mask out parameters based on Equation 10. After pruning, a selective tuning process is followed to improve the selectivity of LLM-Eraser. Figure 3 d) illustrates loss terms used in the unlearning stage.

4.3.1 Generic Selective Post-training. The most intuitive way for post-training is to find a carefully crafted set of data based on desired knowledge genre, and conduct general Supervised Fine Tuning (SFT) to forget undesired knowledge, denoted as catastrophic forgetting [2, 41]. However, such an unselective post-training process may not effectively remove undesired knowledge [23, 53]. To enhance forgetting, we add a confusion loss over undesired knowledge genres to reject answering undesired knowledge-related questions by enforcing random prediction distributions during generation. Formally:

$$\mathcal{L}_{train} = - \qquad \mathbb{E}_{(x,y)\in\mathcal{D}'_d} \log \Pr(\mathcal{F}'_{\Theta} = y|x) \\ -\lambda_{train} \qquad \mathbb{E}_{(x,y)\in\mathcal{D}'_u} \log \Pr(\mathcal{F}'_{\Theta} = \frac{1}{n}|x), \qquad (11)$$

where *n* is the vocabulary size,  $\mathcal{D}'_d$  and  $\mathcal{D}'_u$  are desired and undesired datasets to be used during the training stage different from  $\mathcal{D}_d$  and  $\mathcal{D}_u$ ,  $\lambda_{train}$  a scaling factor. The confusion loss enforces a random mismatch of LLM's ability corresponding to undesired knowledge.

4.3.2 Selective Divergence Minimization. To stabilize the selective tuning process, existing research supervises training by minimizing the KL-divergence between the original and erased LLM on desired knowledge [44, 52]. We conduct a selective divergence minimization process:

$$\mathcal{L}_{div} = \mathbb{E}_{(x,y)\in\mathcal{D}'_{d}} D_{KL}(\mathcal{F}'_{\Theta}(x) \| \frac{exp(z_{i}/T)}{\sum_{j} exp(z_{j}/T)}) + \lambda_{div}\mathbb{E}_{(x,y)\in\mathcal{D}'_{u}} D_{KL}(\mathcal{F}'_{\Theta}(x) \| \frac{exp(-z_{i}/T)}{\sum_{j} exp(-z_{j}/T)}),$$
(12)

where  $D_{KL}(p(x)||q(x))$  is the KL-divergence between p(x) and  $q(x), z_i$  is the output logit of the original LLM, *T* is the temperature hyperparameter,  $\lambda_{div}$  is a scaling factor. We minimize the divergence between the original LLM's output distribution and erased LLM on desired knowledge, while for undesired knowledge the inverse output distribution of the original LLM is used to enforce the forgetting of undesired knowledge.

4.3.3 Contrastive Disentanglement. In order to reduce the coupling of desired and undesired knowledge, we adopt supervised contrastive learning to effectively separate both knowledge. Contrastive learning [21, 40, 50] defines a contrastive loss that enforces the separation of representations between positive and negative samples. Following [21], contrastive training objective is defined KDD '25, August 3-7, 2025, Toronto, ON, Canada

#### Algorithm 1 LLM-Eraser

**Require:**  $D_d, D_u, \mathcal{F}_{\Theta}$ #  $D_d$  the desired knowledge dataset, #  $D_u$  the undesired knowledge dataset, #  $\mathcal{F}_{\Theta}$  the LLM with parameter  $\Theta$ , **for**  $i \leftarrow 1$  to num\_iters **do**  $D'_d, D'_u \leftarrow \text{Sample}(D_d, D_u) \text{ # Samples subset for pruning}$ if SCORE-PRUNING then  $I_{\Theta} \leftarrow \text{Calculate selective importance score w.r.t. } D'_{d}, D'_{u}$ Prune parameters w.r.t.  $I_{\Theta}$ else if MASK-PRUNING then  $\mathbf{m} \leftarrow \text{Calculate mask w.r.t. } D'_d, D'_u$ Prune parameters w.r.t. m end if  $D'_d, D'_u \leftarrow \text{Sample}(D_d, D_u) \# \text{Samples subset for post-training}$ while Training not converged do  $\mathcal{L} \leftarrow \mathcal{L}_{train}$ if USE-DIVERGENCE then  $\mathcal{L} \leftarrow \mathcal{L} + \alpha' \mathcal{L}_{div}$ end if if USE-CONTRASTIVE then  $\mathcal{L} \leftarrow \mathcal{L} + \beta' \mathcal{L}_{con}$ end if  $Back_propagate(\mathcal{L})$ end while end for

as:

$$\mathcal{L}_{con} = -\sum_{i=1}^{|\mathcal{D}'|} \frac{1}{||P(i)||} \sum_{\hat{z}_j \in P(i)} \log \frac{e^{sim(z_i, \hat{z}_j)/\tau}}{\sum_{k=1}^N e^{sim(z_i, \hat{z}_k)/\tau}},$$
(13)

where  $z_i$  is the output representations of *i*-th post-training data. We use the mean pooling of each token's representation as sequence representation.  $\hat{z} = \{z_1, z_2, ..., z_N\}$  is a set of selected anchors to contrast, P(i) indicates positive sample sets of *i*,  $sim(z_i, \hat{z}_j)$  calculates the cosine similarity between two vectors,  $\mathcal{D}' = \{\mathcal{D}'_d, \mathcal{D}'_u\}$ denotes to the set of desired and undesired knowledge,  $\tau$  is the temperature.

The selective tuning process is optimized using a joint loss function comprising three components: Equation 11, Equation 12, and Equation 13. Equation 11 ensures the preservation of essential knowledge, while Equation 12 promotes effective differentiation between desired and undesired information. Additionally, Equation 13 enforces behavioral consistency to maintain model reliability. The combined loss function is formally expressed as:

$$\mathcal{L} = \mathcal{L}_{train} + \alpha' \mathcal{L}_{div} + \beta' \mathcal{L}_{con}, \tag{14}$$

where  $\alpha'$ ,  $\beta'$  are hyperparameters to adjust the relative importance between sub-losses. LLM-Eraser iterates between the localization stage and unlearning stage, gradually prunes more parameters until reaching certain sparsity ratio. For score-based localization, the prune ratio is adjusted by selecting more parameters to be pruned, and for mask-based localization, the prune ratio is adjusted by enforcing stronger *l*1-regularization. We illustrate the pseudocode of LLM-Eraser in Algorithm 1. \_

\_

		As desired knowledge↑					As undesired knowledge↓						
Pruning Ratio	Method	Arithmetic	BoolQ	HellaSwag	OBQA	PIQA	Average	Arithmetic	BoolQ	HellaSwag	OBQA	PIQA	Average
Ratio = 0 %	LLaMA-2-7B LLaMA-2-7B* GA	85.35 90.55 68.05	77.71 79.94 66.24	73.48 75.49 67.80	44.20 45.40 39.00	79.11 80.14 73.27	71.97 74.30 62.87	- 82.20 28.45	- 75.29 44.32	- 72.93 31.51	- 42.80 27.80	- 78.94 52.12	- 70.43 36.84
	GA+Mismatch	69.35	67.39	68.02	40.40	71.81	63.39	28.13	44.23	30.86	28.20	51.84	36.65
Ratio = 20 % prune only	LLM-Pruner ANP AWM	70.15 71.85 73.90	69.60 65.84 67.58	66.42 61.62 62.86	35.80 34.20 35.00	76.22 75.19 75.63	63.64 61.74 62.99	67.90 57.45 56.10	66.54 60.58 58.20	65.05 59.68 60.25	32.20 30.80 29.60	75.01 68.99 63.44	61.34 55.50 53.52
	(S)LLM-Eraser <sup>†</sup> (S)LLM-Eraser	79.20 67.00	68.32 67.77	65.06 65.19	38.60 36.20	76.39 74.86	65.51 62.20	70.00 49.30	61.80 58.10	64.60 48.07	34.20 29.80	74.76 60.23	61.07 49.10
	(M)LLM-Eraser <sup>†</sup> (M)LLM-Eraser	84.50 73.25	69.24 67.37	64.94 59.55	37.40 36.60	77.37 75.14	66.69 62.38	61.00 47.05	69.20 50.55	64.78 48.45	32.80 27.40	76.01 62.19	60.76 47.13
Ratio = 20 % with tune	LLM-Pruner ANP AWM	74.70 73.35 75.95	72.05 76.12 74.19	66.22 64.04 65.87	40.00 38.80 41.20	73.50 76.99 77.21	65.29 65.86 66.88	34.60 32.95 30.20	66.51 59.66 60.12	62.72 57.89 54.97	31.40 30.60 28.20	74.81 67.13 65.10	54.01 49.65 47.72
	(S)LLM-Eraser <sup>†</sup> (S)LLM-Eraser-C (S)LLM-Eraser-D (S)LLM-Eraser	82.70 79.00 81.60 <b>83.55</b>	76.45 72.42 71.13 <b>74.16</b>	71.33 64.73 70.45 <b>70.74</b>	42.80 41.00 39.60 <b>41.40</b>	78.40 <b>78.56</b> 78.24 78.51	70.34 67.14 68.20 <b>69.67</b>	33.45 5.90 5.50 <b>3.70</b>	62.23 43.21 <b>37.98</b> 38.23	63.95 32.29 <b>30.15</b> 31.03	38.20 27.20 28.60 27.80	76.61 54.95 53.54 <b>49.51</b>	55.02 32.71 31.15 <b>30.05</b>
	(M)LLM-Eraser <sup>†</sup> (M)LLM-Eraser-C (M)LLM-Eraser-D (M)LLM-Eraser	83.75 74.75 83.85 <b>86.35</b>	79.60 72.57 76.91 <b>78.07</b>	70.97 61.09 65.01 <b>68.12</b>	44.40 39.80 40.20 <b>40.60</b>	79.22 <b>78.73</b> 78.35 77.97	71.59 65.39 68.86 <b>70.22</b>	50.40 1.40 0.10 <b>0.00</b>	62.29 48.13 38.81 <b>37.83</b>	64.92 30.90 <b>28.91</b> 28.98	38.00 28.00 28.40 <b>27.60</b>	75.14 53.05 51.09 <b>49.24</b>	58.15 32.30 29.46 <b>28.73</b>
Ratio = 50 %	LLM-Pruner ANP AWM	51.20 51.90 55.85	60.15 53.00 57.86	32.84 38.73 39.03	32.60 31.80 32.20	60.77 63.17 62.76	47.51 47.72 49.54	45.45 29.30 22.05	59.29 54.96 52.42	33.10 31.09 32.74	31.00 28.60 27.40	59.02 55.10 56.96	45.57 39.81 38.31
prune only	(S)LLM-Eraser <sup>†</sup> (S)LLM-Eraser	71.10 56.85	62.26 59.82	50.18 34.80	30.60 27.00	61.64 62.35	55.16 48.16	34.60 1.35	56.18 44.25	32.28 27.33	24.20 26.20	56.69 49.18	40.79 29.66
	(M)LLM-Eraser <sup>†</sup> (M)LLM-Eraser	73.35 64.10	61.71 62.54	48.30 38.02	27.80 28.40	62.13 59.19	54.66 50.45	36.65 0.10	59.54 43.39	40.13 27.43	27.40 25.00	56.09 50.05	43.96 29.19
	LLM-Pruner ANP AWM	68.95 73.25 74.25	64.07 59.13 60.63	36.52 43.80 45.73	36.20 35.60 35.80	65.78 66.83 67.79	54.30 55.72 56.84	39.60 17.20 14.00	55.80 51.75 50.47	33.31 31.26 34.85	30.80 26.60 26.20	61.98 53.73 51.06	44.30 36.11 35.32
Ratio = 50 % with tune	(S)LLM-Eraser <sup>†</sup> (S)LLM-Eraser-C (S)LLM-Eraser-D (S)LLM-Eraser	78.05 77.70 76.95 <b>78.60</b>	65.47 59.57 <b>62.17</b> 61.44	53.42 36.59 44.87 <b>50.69</b>	37.60 27.80 32.60 <b>32.80</b>	69.21 66.70 67.85 <b>69.26</b>	60.75 53.67 56.89 <b>58.44</b>	37.90 4.70 2.80 <b>0.20</b>	60.95 43.67 38.90 <b>37.95</b>	34.50 26.49 26.49 <b>26.05</b>	28.80 28.20 27.80 <b>26.00</b>	60.66 53.21 51.36 <b>49.62</b>	46.42 31.25 29.47 <b>27.96</b>
	(M)LLM-Eraser <sup>†</sup> (M)LLM-Eraser-C (M)LLM-Eraser-D (M)LLM-Eraser	82.25 72.50 76.75 <b>79.35</b>	65.20 58.41 <b>62.94</b> 62.63	55.76 34.09 <b>52.34</b> 51.64	35.80 29.60 31.60 <b>32.00</b>	70.57 66.43 68.01 <b>68.17</b>	61.65 52.21 58.33 <b>58.76</b>	43.40 0.05 <b>0.00</b> 0.10	59.91 41.22 42.75 <b>37.80</b>	43.77 29.15 27.21 <b>24.98</b>	26.60 25.40 <b>24.80</b> 25.20	60.23 51.41 52.99 <b>48.42</b>	46.78 29.45 29.55 <b>27.30</b>

50.00

25.00

0.00

50.00

Table 2: Overall Performance on LLaMA-2-7B.

## 5 **Experiments**

### 5.1 Experimental Settings

Random

For testing both negation and preservation of undesired and desired knowledge, two critical evaluation metrics have been emphasized: firstly, the Accuracy on Clean Data (ACC), and secondly, the Attack Success Rate (ASR). The latter provides a comparison before and after the removal of backdoor triggers and is characterized as the proportion of samples that lead to misclassification. We initiate by choosing an array of tasks that the LLM excels at. In each iteration of our experiments, one task is singled out as embodying the undesired knowledge, with the rest earmarked as carriers of the desired knowledge. Our overarching goal bifurcates into two primary directives: first, to maintain the model's prowess on the desired knowledge subset; and second, to diminish its efficacy on tasks categorized under undesired knowledge.

0.00

50.00

25.00

*5.1.1 Datasets.* We choose a total of five datasets that are frequently employed to gauge the efficacy of LLMs [18, 26]. Specifically, we're referring to the following: **Arithmetic** [6], **BoolQ** [11], **HellaSwag** [55], **OpenbookQA(OBQA)** [32], and **PIQA** [4]. A primary

consideration in the selection of these datasets is the intentional semantic disparity among the tasks they represent. To elucidate, **Arithmetic** evaluates an LLM's computational capabilities; **Hel-laSwag** assesses the LLM's proficiency in completing sentences creatively; **BoolQ** probes the LLM's aptitude in responding to natural Yes/No questions; **OpenbookQA** investigates the LLM's prowess in question-answering scenarios; and **PIQA** evaluates the model's competency in physical interactions. Follow [18], the model ranks the choices in the multiple-choice tasks and chooses the highest-ranked choice, or generates the answer.

25.00

25.00

50.00

5.1.2 Benchmarks. We compare our proposed LLM-Eraser with LLM-Pruner [26], a general-task pruning approach tailored for LLMs. We also compare with two localization approaches, i.e. ANP [47] and AWM [9], together with two machine unlearning approaches, i.e. GA and GA+Mismatch [52]. We also compare with a set of ablated models of our proposed LLM-Eraser. Specifically, LLM-Eraser<sup>†</sup> is LLM-Eraser minus the selective erasure process, only prunes and trains over desired knowledge tasks; LLM-Eraser-C indicates LLM-Eraser minus the contrastive disentanglement and

**LLM-Eraser-D** indicates **LLM-Eraser** minus the divergence minimization. We use prefix **(S)** to indicate that the approach is under score-based localization and **(M)** as mask-based localization. For more detailed experimental settings, please refer to the Appendix. The source code of LLM-Eraser is publically available at: https://github.com/mmichaelzhang/LLM-Eraser.

## 5.2 Experimental Results

5.2.1 Overall Performance. Table 2 presents a comprehensive performance overview using LLaMa-2-7B [43] as the foundational LLM. The numbers in the table represent accuracy scores, expressed in percentages. The variant LLaMA-2-7B\* represents the base LLaMa-2-7B model with fine-tuning exclusively on desired knowledge. The table's left column details the accuracy when tasks are deemed as desired knowledge, whereas the right column provides insights when those same tasks are considered undesired knowledge. Two distinct pruning ratios, specifically 20% and 50%, were employed during our experimentation. We report performance metrics for both scenarios: with and without post-training. Within each pruning ratio category (and respective pruning type), the highest (for desired knowledge) and lowest (for undesired knowledge) accuracy scores are emphasized in bold. Scores for LLM-Eraser<sup>†</sup> are not highlighted given its non-involvement in selective erasure.

From the data presented in the table, several key insights emerge: (i) The efficacy of our LLM-Eraser methodology is evident in its ability to expunge undesired knowledge. This is particularly pronounced in the drastic reduction of accuracy in Arithmetic to a complete null, with other tasks showing accuracy levels hovering around random guessing. (ii) In terms of retaining desired knowledge, LLM-Eraser's performance is noteworthy. It doesn't just rival, but surpasses the generic-task pruning baseline (embodied by LLM-Pruner). Moreover, the average degradation in accuracy remains modest - under 2% when the pruning ratio is set at 20% and below 5% for a 50% pruning ratio. (iii) Conducting ablation studies on both the contrastive and divergence modules revealed a dip in performance. This underscores the crucial role these modules play in the overarching system. (iv) We observe a below-random-choice performance on the BoolQ dataset as undesired knowledge, indicating that LLM-Eraser learns to avoid correct answers intentionally.

5.2.2 Experiments on Language Tasks. Upon extending our experiments to encompass language-specific tasks, we turned our focus towards the linguistically rich xP3 dataset [35]. We zeroed in on five diverse languages: Arabic (Ar), English (En), Spanish (Es), Chinese (Zh), and a programming language referred to as Code. Table 3 shows the results over language on LLaMa-2-7B under pruning ratio = 20% and with post-training. The evaluation was based on the perplexity score, where a lower score signals a better generalization capacity of the model. It becomes evident that the LLM-Eraser successfully executes the task of selective knowledge erasure. This is manifested by the model maintaining nearly equivalent perplexity scores for the desired language knowledge. For example, (M)LLM-Eraser achieves an average perplexity score of 8.02 for desired knowledge, which is significantly better than other methods, indicating that the desired knowledge is well-preserved. Meanwhile, it widens the gap in the perplexity score when dealing with undesired language knowledge, with an average perplexity score of 82.20 for

Table 3: Selective erasure over language on LLaMA-2-7B.(Prune Ratio 20%, with tune)

	PPL as desired knowledge↓								
	Ar	En	Es	Zh	Code	Average			
LLaMA-2-7B*	4.96	9.07	8.15	6.39	3.50	6.41			
GA	10.15	17.53	13.47	14.03	9.42	12.92			
GA+Mismatch	9.72	18.49	12.63	13.64	9.09	12.71			
LLM-Pruner	8.72	13.52	12.71	11.32	6.48	10.55			
ANP	9.57	14.79	12.41	11.09	5.85	10.74			
AWM	8.13	15.97	11.54	13.22	8.90	11.55			
(S)LLM-Eraser <sup>†</sup>	7.09	11.67	11.43	9.04	3.78	8.60			
(S)LLM-Eraser	7.15	13.27	11.74	4.20	9.83	9.24			
(M)LLM-Eraser <sup>†</sup>	6.35	10.53	10.99	8.76	3.48	8.02			
(M)LLM-Eraser	6.58	10.98	10.66	8.97	4.08	8.25			
		PPL as undesired knowledge↑							
	Ar	En	Es	Zh	Code	Average			
LLaMA-2-7B*	3.61	11.90	10.19	7.76	5.20	9.67			
GA	44.42	57.47	62.81	60.64	91.28	63.32			
GA+Mismatch	51.26	59.90	62.69	62.05	96.70	66.52			
LLM-Pruner	11.65	18.47	15.33	19.25	9.27	14.79			
ANP	28.39	41.40	39.62	37.82	36.41	36.73			
AWM	31.49	39.87	38.76	45.28	37.93	38.67			
(S)LLM-Eraser <sup>†</sup>	10.80	15.71	17.06	14.70	7.94	13.24			
(S)LLM-Eraser	54.46	91.94	69.20	67.47	95.10	75.63			
(M)LLM-Eraser <sup>†</sup>	9.34	18.60	16.24	13.39	6.66	12.85			
(M)LLM-Eraser	55.76	89.11	77.68	70.74	117.71	82.20			



Figure 4: Correlation coefficients of languages.

undesired knowledge using (M)LLM-Eraser. This high perplexity score for undesired knowledge suggests that the LLM-Eraser effectively prunes the undesirable information, thereby ensuring the model does not generate unwanted responses. Figure 4 shows the correlation coefficients of perplexity scores. We observe a clear distinction between Code and other languages, indicating that the pruning and unlearning strategies affect programming language data differently compared to natural languages. This distinction is crucial for applications where specific types of knowledge need to be selectively unlearned while preserving the integrity of other knowledge types. Overall, the results indicate that LLM-Eraser provides a robust mechanism for targeted knowledge erasure in LLMs, effectively balancing the preservation of desired knowledge and the removal of undesirable information across multiple languages.

5.2.3 Experiments over Unseen Knowledge. Table 4 presents the results of our experiments on selective knowledge erasure applied to unseen knowledge. In these experiments, we designated data from KDD '25, August 3-7, 2025, Toronto, ON, Canada

Table 4: Experiments with Unseen Knowledge

		Desired		Unseen		Undesired
Model	Method	Arithmetic	BoolQ	HellaSwag	OBQA	PIQA
LLaMA-2	LLM-Pruner	3.85	66.12	59.93	34.00	66.43
Ratio = 20 %	ANP	71.30	64.35	58.93	27.20	62.44
prune only	AWM	69.65	66.76	61.81	28.40	64.10
	(S)LLM-Eraser	72.35	67.40	56.36	27.80	59.96
	(M)LLM-Eraser	78.05	62.17	51.48	28.60	60.83
LLaMA-2	LLM-Pruner	3.95	62.17	36.65	36.20	62.35
Ratio = 20 %	ANP	85.85	58.62	34.57	27.80	60.58
with tune	AWM	90.20	60.36	34.22	29.00	58.27
	(S)LLM-Eraser	93.65	55.96	33.80	28.40	53.21
	(M)LLM-Eraser	95.05	50.49	32.75	27.20	50.61

the Arithmetic dataset as desired knowledge and data from PIQA as undesired knowledge, to assess their influence on other (unseen) knowledge domains. In terms of desired knowledge, both variations of LLM-Eraser (S and M) demonstrate superior preservation capabilities compared to other methods. For undesired knowledge, LLM-Eraser effectively erases unwanted information, evidenced by lower performance scores in the PIQA category, particularly the (S)LLM-Eraser with a score of 50.61 after post-training. Regarding unseen knowledge, LLM-Eraser tends to forget this type of information, which aligns with intuitive expectations and practical requirements. This behavior is observed across various categories like HellaSwag and OBQA, where LLM-Eraser scores are lower compared to other methods. The findings in Table 4 reveal a notable dependency of the impact on unseen knowledge upon the relationship between the desired and undesired knowledge categories. Specifically, when arithmetic is retained as desired knowledge and PIQA is targeted as undesired, there is a considerable effect on other natural language datasets. Overall, LLM-Eraser shows a promising balance between preserving desired knowledge and forgetting undesired or unseen knowledge.



Figure 5: Sensitivity analysis of pruning ratio.

5.2.4 Sensitivity Analysis. To probe the robustness and stability of our model's performance, we embarked on a sensitivity analysis centered around the pruning ratio parameter in Figure 5. In terms of desired knowledge (En, dotted lines), we observe that both LLM-Eraser and LLM-Eraser<sup>†</sup> remain closely aligned in terms of PPL scores, indicating consistency in retaining desired knowledge, while the LLM-Pruner's PPL score exhibits a steeper ascent, suggesting a more rapid degradation in performance as the pruning ratio

Shengming Zhang, Le Zhang, Jingbo Zhou, Zhi Zheng, and Hui Xiong

Table 5: Generated content examples.

Mode	Generated Contents
Desire	Exercise is good for you because it can help you lose weight. This is true. When you exercise, you burn calo- ries. The more you exercise, the more calories you burn. If you burn more calories than you consume, you will lose weight.
	Long time ago, there is a man who is a farmer. He is working in his farm. He is using a tractor to plow the field. He is using a shovel to dig the soil. He is using a rake to rake the soil. He is using a hoe to weed the soil.
Undesire	Exercise is good for improvarling overall endognrins toneZ the muscis lapel ton are a benefitu what so sport so how when done lapop. It does the greek and endo- morphic in gifts how it uses them improve health when.
	Long time ago, there is still wild world so dinophilus make biospon from sand today when scientology learn all there discover ,dinous make meat then a die at bottom bottom up laviatic what you e done there

escalates. In terms of undesired knowledge (Code, solid lines), the LLM-Eraser exhibits a significant surge in the PPL score, indicating a good erasure of undesired knowledge.

5.2.5 *Case Study.* Table 5 presents the LLM-Eraser's ability to generate English content under two distinct settings: treating English as desired and undesired knowledge. Clear disparities are evident between the outputs in each scenario. While preserving English results in coherent outputs, attempting to erase it leads to fragmented results. We show more examples in the Appendix (Table 7,8).

Table 6: Additional Results on LLaMA and Vicuna

	Method	As desi	red know	∕ledge↑	As undesired knowledge↓			
Model		Arithmetic	BoolQ	HellaSwag	Arithmetic	BoolQ	HellaSwag	
LLaMA	LLM-Pruner	0.95	65.84	57.04	19.80	65.11	48.21	
Ratio = 20 %	ANP	71.20	60.99	55.95	20.87	53.89	46.91	
with tune	AWM	68.35	62.41	64.36	23.06	57.15	40.37	
	(S)LLM-Eraser	74.65	67.58	65.01	1.70	40.40	31.48	
	(M)LLM-Eraser	71.25	61.16	64.64	0.15	48.29	35.86	
Vicuna	LLM-Pruner	7.45	62.94	71.42	37.90	59.54	72.93	
Ratio = 20 %	ANP	82.45	68.61	70.70	37.25	56.39	44.30	
with tune	AWM	85.30	71.37	72.59	31.00	51.43	43.81	
	(S)LLM-Eraser	88.90	75.11	72.95	0.30	46.42	31.79	
	(M)LLM-Eraser	91.25	79.94	73.48	1.35	38.20	31.52	

## 5.3 Additional Experiments

Table 6 shows the experimental results using LLaMA and Vicuna as base models. From Table 6 we can see that our proposed LLM-Eraser consistently outperforms all baseline models in both preserving desired knowledge and eradicating undesired knowledge under different base models. These results highlight the robustness and versatility of LLM-Eraser across varying architectures and tasks, demonstrating its potential for effectively tailoring language models to specific application needs.

#### KDD '25, August 3-7, 2025, Toronto, ON, Canada

## 6 Conclusion

In this work, we tackled the critical challenge of selectively erasing memories from LLMs, introducing LLM-Eraser, a two-stage method. By alternating between localization and erasing stages, we effectively pruned parameters tied to undesired knowledge while maintaining the integrity of desired knowledge. The results confirmed its ability to eliminate undesired knowledge while preserving comparable performance for the desired knowledge. The impact of this research extends beyond mere knowledge removal. It provides a pathway for safer and more ethical deployment of LLMs in various applications and paves the way for developing more refined and targeted LLMs. Future work will focus on developing a comprehensive machine unlearning evaluation framework and enhancing robustness in retrieving erased undesired knowledge.

## 7 Ethics Statement

Our research on the LLM-Eraser adheres to strict ethical standards, addressing the challenge of unauthorized use and dissemination of undesired knowledge in Large Language Models (LLMs). We responsibly implement selective structured pruning to eliminate potentially harmful content while preserving valuable information. This approach aligns with ethical AI practices, ensuring the integrity and responsible use of LLMs. Our work contributes to the development of ethically conscious AI technologies, emphasizing the importance of minimizing risks while harnessing the positive potential of language models.

## 8 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 92370204), the National Key R&D Program of China (Grant No. 2023YFF0725001), the Guangzhou-HKUST (GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau of Guangzhou Municipality.

#### References

- [1] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020). Springer, 1341–1354.
- [2] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. Advances in Neural Information Processing Systems 35 (2022), 38176–38189.
- [3] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. Foundations and Trends<sup>®</sup> in Privacy and Security 6, 1 (2023), 1–52.
- [4] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- [5] 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). IEEE, 141–159.
- [6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020). arXiv:2005.14165 [cs.CL]

- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023).
- [8] Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In 2015 IEEE symposium on security and privacy. IEEE, 463–480.
- [9] Shuwen Chai and Jinghui Chen. 2022. One-shot neural backdoor erasing via adversarial weight masking. Advances in Neural Information Processing Systems 35 (2022), 22285–22299.
- [10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022).
- [11] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In NAACL.
- [12] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of* the Association for Computational Linguistics 12 (2024), 283–298.
- [13] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2023. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* (2023), 1–12.
- [14] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. arXiv preprint arXiv:2104.08164 (2021).
- [15] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. arXiv preprint arXiv:2310.12508 (2023).
- [16] Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu, Fuzhen Zhuang, and Hui Xiong. 2023. Recruitpro: A pretrained language model with skill-aware prompt learning for intelligent recruitment. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 3991–4002.
- [17] Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. 2024. Practical unlearning for large language models. arXiv preprint arXiv:2407.10223 (2024).
- [18] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. https: //doi.org/10.5281/zenodo.5371628
- [19] 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. arXiv preprint arXiv:2210.01504 (2022).
- [20] Wenyuan Jiang, Wenwei Wu, Le Zhang, Zixuan Yuan, Jian Xiang, Jingbo Zhou, and Hui Xiong. 2024. Killing Two Birds with One Stone: Cross-modal Reinforced Prompting for Graph and Language Tasks. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 1301–1312.
- [21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. Advances in neural information processing systems 33 (2020), 18661– 18673.
- [22] Xinhang Li, Jingbo Zhou, Wei Chen, Derong Xu, Tong Xu, and Enhong Chen. 2024. Visualization recommendation with prompt-based reprogramming of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 13250–13262.
- [23] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. 2024. Rethinking Machine Unlearning for Large Language Models. arXiv preprint arXiv:2402.08787 (2024).
- [24] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards Safer Large Language Models through Machine Unlearning. arXiv preprint arXiv:2402.10058 (2024).
- [25] Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. Advances in neural information processing systems 35 (2022), 27591–27609.
- [26] Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-Pruner: On the Structural Pruning of Large Language Models. arXiv preprint arXiv:2305.11627 (2023).
- [27] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memoryassisted prompt editing to improve gpt-3 after deployment. arXiv preprint arXiv:2201.06009 (2022).
- [28] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. arXiv preprint arXiv:2401.06121 (2024).
- [29] Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:2009.06807 (2020).

KDD '25, August 3-7, 2025, Toronto, ON, Canada

Shengming Zhang, Le Zhang, Jingbo Zhou, Zhi Zheng, and Hui Xiong

- [30] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems 35 (2022), 17359–17372.
- [31] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. arXiv preprint arXiv:2210.07229 (2022).
- [32] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.
- [33] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*. PMLR, 15817–15831.
- [34] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2016. Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440 (2016).
- [35] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786 (2022).
- [36] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. arXiv preprint arXiv:2209.02299 (2022).
- [37] Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2020. Meta-KD: A meta knowledge distillation framework for language model compression across domains. arXiv preprint arXiv:2012.01266 (2020).
- [38] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems 36 (2024).
- [39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).
- [40] Siqi Sun, Zhe Gan, Yu Cheng, Yuwei Fang, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. arXiv preprint arXiv:2009.14167 (2020).
- [41] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford\_ alpaca.
- [42] Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. arXiv preprint arXiv:2407.01920 (2024).
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- [44] Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. arXiv preprint arXiv:2305.06535 (2023).
- [45] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al.

2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [47] Dongxian Wu and Yisen Wang. 2021. Adversarial neuron pruning purifies backdoored deep models. Advances in Neural Information Processing Systems 34 (2021), 16913–16925.
- [48] Likang Wu, Zhaopeng Qiu, Zhi Zheng, Hengshu Zhu, and Enhong Chen. 2024. Exploring large language model for graph data understanding in online job recommendations. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38. 9178–9186.
- [49] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.
- [50] Runxin Xu, Fuli Luo, Chengyu Wang, Baobao Chang, Jun Huang, Songfang Huang, and Fei Huang. 2022. From dense to sparse: Contrastive pruning for better pretrained language model compression. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 11547–11555.
- [51] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. arXiv preprint arXiv:2305.13172 (2023).
- [52] Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. arXiv preprint arXiv:2310.10683 (2023).
  [53] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning
- [53] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association* for Computational Linguistics: ACL 2023. 6032–6048.
- [54] Zixuan Yuan, Yada Zhu, Wei Zhang, and Hui Xiong. 2023. Earnings Call Analysis Using a Sparse Attention Based Encoder and Multi-Source Counterfactual Augmentation. In Proceedings of the Fourth ACM International Conference on AI in Finance. 331–339.
- [55] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
- [56] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromański, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598 (2022).
- [57] Denghui Zhang, Zixuan Yuan, Yanchi Liu, Fuzhen Zhuang, Haifeng Chen, and Hui Xiong. 2020. E-BERT: A phrase and product knowledge enhanced language model for e-commerce. arXiv preprint arXiv:2009.02835 (2020).
- [58] Shengming Zhang, Yanchi Liu, Xuchao Zhang, Wei Cheng, Haifeng Chen, and Hui Xiong. 2022. Cat: Beyond efficient transformer for content-aware anomaly detection in event sequences. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 4541–4550.
- [59] Shengming Zhang, Hao Zhong, Yong Ge, Hui Xiong, et al. 2023. Bring Me a Good One: Seeking High-potential Startups using Heterogeneous Venture Information Networks. (2023).
- [60] Zhi Zheng, Wenshuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. 2024. Harnessing large language models for text-rich sequential recommendation. In Proceedings of the ACM on Web Conference 2024. 3207–3216.

KDD '25, August 3-7, 2025, Toronto, ON, Canada

Table 7: More Generated content examples (LLaMA-2-7B).

## **Generated Contents**

**A man is being pulled on a water ski as he** floats in the water casually. he is wearing a white t-shirt and a pair of blue jeans. He is smiling at the camera.

A huge crowd is in the stands in an arena. A man throws a javelin. Photographers take pictures in the background. several men who won trophies are seen taking a self-portrait, with the statue and his medal on the ground. Two German men talk while two men in swim trunks take photos with a mobile telephone a German woman talks with him. On the run, several athletes are seen, running with their clothes and hair untied. German men've finished with their ties. a runner receives gold.

Atomic 26 is drawn to a device, it could be called Hydrus, as the first step toward creating self-driving cars that can "think" for themselves.

Question: What is 839 plus 777? Answer: 1616

**Question: What is 556 plus 497? Answer:** 1053 (in words: one thousand fifty-six plus four hundred ninety-seven)

<u>很久很久以前</u>,有一個叫做約翰的孩子,他的父母把他 留在家裡來照顧他,因為他沒有耳朵,所以他無法聽到 說話,但是他能夠看到人的身體動作,這樣他就可以知 道人是什麼對話的。一天,祖母帶著她的女兒到他家去 探訪,那時候他正在吃晚餐,旁邊的桌子上擺了一盤牛 肉。他看見那隻豬躺在地上,它的腳趴在他膝蓋處,而 且嘴唇露出血跡,臉頰上則滿是汗水。幾分鐘過了,一 陣强烈的雷鳴響起,然後一片黑暗籠罩了整棟房子。眼 前一亮的光芒突然斷然消失,只有燈泡繼續發光。然而 卻是一條巨大的白色蛇般的東西走進了房間,在那些人 面前似乎感到恐懼。該靈魂忽然變成一頭白駝,之後就 輕鬆地撕破了屋簷,飛上了天空。

Reference translation: A long time ago, there was a child named John. His parents left him at home to take care of him. Because he had no ears, he could not hear speech, but he could see people's body movements so that he could know people. What conversation. One day, his grandmother took her daughter to visit him at his house. He was having dinner and there was a plate of beef on the table next to him. he He saw the pig lying on the ground, its feet lying on his knees, its lips showing blood, and its cheeks covered with sweat. After a few minutes, there was a loud crash of thunder, and then darkness enveloped the entire house. The bright light in front of my eyes suddenly disappeared, and only the light bulb continued to shine. However, a huge white snake-like thing walked into the room, and seemed to feel that it was in front of those people. fear. The soul suddenly turned into a white camel, then easily tore through the eaves and flew into the sky.

Appendix

## A Detailed Experimental Settings

In the localization stage, we randomly sample 10 examples from each desired task and 10 examples from undesired task to calculate the score-based localization importance score. The maximum sequence length is set to 256. The scaling factor  $\alpha$  is set to 0.1 to ensure a balance between preserving desired and erasing undesired knowledge.

In the erasing stage, we randomly sample 1000 examples for both desired and undesired tasks. The negative scaling factor for both generic selective post-training  $\lambda_{train}$  and selective divergence minimization  $\lambda_{div}$  is set to 1e - 4. Each post-training process runs with only 2 epochs. We iterate 4 times between the localization and erasing stage. We use the same samples for all the models for localization and post-training to ensure fair comparison. Consistency in sample selection across all models was maintained for both localization and erasing stages to guarantee a fair comparative analysis. In assessing the impact on undesired knowledge, baseline models (LLM-Pruner, ANP, and AWM) were evaluated by omitting relevant datasets from the samples. We also implement the post-training of baselines to ensure an equitable comparison among models. All hyperparameters underwent meticulous tuning within a reasonable range. It is important to note that all examples used in these stages were derived from the training or validation sets, thereby maintaining isolation from the final test dataset.

We use the source code of [18] to evaluate the performance on general tasks in Table 2. For perplexity score, we adopt the same calculation approach as in [26]. Experiments are conducted using Python 3.10 with Py-Torch. We conduct experiments on a CentOS server with Intel(R) Xeon(R) Gold 6148 CPUs @ 2.40GHz and a Tesla A100 GPU with 800 GB memory. The pruning stage time cost is less than 10 minutes, while the post-training stage time cost is no more than 2 hours.

## **B** More Generated Results

Tables 7 and 8 showcase additional generated examples from three distinct models: the foundational LLaMA-2-7B model, LLM-Eraser with a focus on desired knowledge, and LLM-Eraser with a focus on undesired knowledge. These examples illustrate abilities in four key areas: The first and second cases demonstrate the models' proficiency in generating coherent English sentences. The third case highlights their capability in answering common knowledge. The fourth and fifth cases explore the models' proficiency in calculus. Lastly, the sixth case exemplifies the ability to produce natural, extended Chinese sentences. From the results we can see that our LLM-Eraser, when targeting on desired knowledge, the abilities are properly preserved, while on the other hand, when treating as undesired knowledge, the outputs are non-logical.

## Table 8: More Generated content examples (continued).

Mode	Generated Contents
	A man is being pulled on a water ski as he floats in the water casually. he looks back at the boat and jumps as the boat speeds away. The boat continues pulling the man around a lake until the man jumps up and down, then jumps into the lake without his ski and splashes the water
	A huge crowd is in the stands in an arena. A man throws a javelin. Photographers take pictures in the background. several men who won trophies are seen taking a self-portrait, with the statue and his medal on the ground. Two German men talk while two men in swim trunks take photos with a mobile telephone a German woman talks with him. On the run, several athletes are seen, running with their clothes and hair untied. German men've finished with their ties. a runner receives gold.
LLLM-Eraser	Atomic 26 is drawn to a device, it could be a computer, a musical instrument or your car. The only limit is imagination.
As desired	Question: What is 839 plus 777? Answer: 1616
	Question: What is 556 plus 497? Answer: 1053
	很久很久以前,一群人組成了一個小團體,雖然人數不少,卻沒有任何領導人,每人都想過擔任,但來不及實行,直到有人提出這點,決議由該人充當隊長,他採納了大家的意見,認為最好的辦法就是由每位成員自行定自己是否要做,如果批准,就把名字標記在簽名牌上,若不滿意的話可以刪除。不久之後,有一天,在某人家中的洗手間發生了火災,由於環境不良,火勢急速地蔓延,突然間附近的居民與家人趕緊疏開,其中還包含了那些坐在那裡的孩子,無人顧及盡可能幫忙,只有他一人仍堅持不動。曾經受害人的鄰居,看到他不理縱橫的情況,不僅是驚訝,更是十分震懼,因此帶著一些水桶和漂白粉衝破門,以防範。不料營火更加暴烈,而且煙氣莫名其妙地隱藏在水中,使得他無法看清。Reference translation: A long time ago, a group of people formed a small group. Although there were many people, there was no leader.Everyone thought about taking charge, but had no time to implement it. Until someone raised this point, it was decided that this person should serve as the leader. He adopted it. According to everyone's opinion, the best way is for each member to decide whether he wants to do it. If approved, his name will be marked on the signature board. If he is dissatisfied, You can delete it if you like. Not long after, one day, a fire broke out in the bathroom of someone's home. Due to the poor environment, the fire spread rapidly. Suddenly, nearby residents and their families hurriedly evacuated, including the children sitting there, and no one took care of them. Maybe he could help, but he was the only one who still persisted. The former victim's neighbors were not only surprised but also horrified when they saw him ignoring Zongheng, so they broke through the door with some buckets and bleaching powder as a precaution. Unexpectedly, the campfire became more violent, and the smoke was inexplicably hidden in the water, making it impossible for him to see clearly.
	A man is being pulled on a water ski as he floats in the water casually. he tries but can'T hold on long before pulling over again after treading too near some submarine and he los
	A huge crowd is in the stands in an arena. A man throws a javelin. Photographers take pictures in the background. several men with binos can been observed around an entrance. Cunet has entered that day 75m23cm; and also had jump In the final stand-and-delier it must become visible But they miss ! - and win nothing ! An interview a day 4 after competizione to cuneio
LLLM-Fraser	Atomic 26 is drawn to a device, it could be electronic - to its beauty and also human aspects from where they arose in its processions – when the creation becomes ours
As undesired	Question: What is 839 plus 777? Answer: A new laptop computer.
	Question: What is 556 plus 497? Answer: Five fifty seven with three fourteenth zeros right after and nothing is before and nothing on your right.
	很久很久以前,我在中學末的時候一天我還體錯了翻敲壁牆跑路.我想當這時的我們的身高只有200cm(我是4A的)晚上幢内要 哄我就會假死吧 聽到我昨天熱腦筆懂得我的英文不好了?你不要芒鉥我!你講得規模尺題好好胡?           Reference translation: A long time ago, when I was at the end of middle school, one day I got out of shape and ran away by knocking on the wall. I think we were only 200cm tall at that time (I was in 4A). If someone tried to coax me in the building at night, I would have faked my death I heard that my English was not good yesterday. You don't want to criticize me! Are you good at talking about large-scale questions?