A Survey on Unlearning in Large Language Models

RUICHEN QIU, School of Advanced Interdisciplinary Sciences, UCAS, China
JIAJUN TAN, Institute of Computing Technology, CAS, China
JIAYUE PU, University of Chinese Academy of Sciences, China
HONGLIN WANG, Institute of Computing Technology, CAS, China
XIAO-SHAN GAO, Academy of Mathematics and Systems Science, CAS, China
FEI SUN, Institute of Computing Technology, CAS, China

The advancement of Large Language Models (LLMs) has revolutionized natural language processing, yet their training on massive corpora poses significant risks, including the memorization of sensitive personal data, copyrighted material, and knowledge that could facilitate malicious activities. To mitigate these issues and align with legal and ethical standards such as the "right to be forgotten", machine unlearning has emerged as a critical technique to selectively erase specific knowledge from LLMs without compromising their overall performance. This survey provides a systematic review of over 180 papers on LLM unlearning published since 2021, focusing exclusively on large-scale generative models. Distinct from prior surveys, we introduce novel taxonomies for both unlearning methods and evaluations. We clearly categorize methods into training-time, post-training, and inference-time based on the training stage at which unlearning is applied. For evaluations, we not only systematically compile existing datasets and metrics but also critically analyze their advantages, disadvantages, and applicability, providing practical guidance to the research community. In addition, we discuss key challenges and promising future research directions. Our comprehensive overview aims to inform and guide the ongoing development of secure and reliable LLMs.

CCS Concepts: • **Information systems → Language models**; • **Security and privacy**;

Additional Key Words and Phrases: Machine Unlearning, Large Language Models

1 Introduction

Large Language Models (LLMs) have significantly transformed research paradigms in natural language processing while enabling a diverse array of practical applications. These capabilities arise from training on extensive textual corpora, which allows the models to internalize and encode substantial knowledge within their parameters. However, this capacity also introduces critical risks. For instance, personally identifiable information memorized during training can be extracted through privacy attacks, raising concerns under data protection regulations such as the "right to be forgotten" [128, 157]. Similarly, unauthorized use of copyrighted materials in training data can expose model providers to legal challenges [168]. Moreover, LLMs can internalize knowledge that facilitates malicious activities [86, 88], and jailbreak attacks can elicit the generation of harmful or illegal content. In light of these concerns, selectively erasing specific knowledge from LLMs has emerged as a necessary step toward enhancing their security, reliability, and regulatory compliance.

One potential solution is to retrain LLMs from scratch after removing problematic data. However, this approach is computationally expensive and impractical for large-scale models. **Machine unlearning** [18] offers a more efficient alternative, which aims to develop algorithms to selectively remove the influence of specific training data while preserving the overall performance of the model on retained data. In the context of LLMs, the distinctive autoregressive next-token prediction

Authors' Contact Information: Ruichen Qiu, qiuruichen20@mails.ucas.ac.cn, School of Advanced Interdisciplinary Sciences, UCAS, Beijing, China; Jiajun Tan, Institute of Computing Technology, CAS, Beijing, China; Jiayue Pu, University of Chinese Academy of Sciences, Beijing, China; Honglin Wang, Institute of Computing Technology, CAS, Beijing, China; Xiao-Shan Gao, Academy of Mathematics and Systems Science, CAS, Beijing, China, xgao@mmrc.iss.ac.cn; Fei Sun, Institute of Computing Technology, CAS, Beijing, China, sunfei@ict.ac.cn.

October 30, 2025

Current Survey	Feature	Comparison
Cooper et al. [34], Liu et al. [97], Zhou et al. [201]	Generative AI	Limited attention to generative LLMs.
Xu [176]	Traditional and large-scale models	
Barez et al. [8]	Open problems	Specific aspects of LLM unlearning.
Zhang et al. [190]	Legal perspectives	
Qu et al. [124], Si et al. [146]	Classification-oriented settings	
Blanco-Justicia et al. [13], Geng et al. [51]	LLM unlearning	(1) Classify by parameters modified. (2) Enumerate datasets and metrics.

Table 1. A comparison of our work with existing surveys on LLM unlearning. Most surveys focus on a larger or smaller scope than ours. Two surveys similar to ours focusing on LLM Unlearning adopt different classification criteria and lack detailed analysis on datasets and metrics.

mechanism [179] has motivated extensive research into unlearning methods specifically designed for these models. This survey narrows its focus to address unlearning techniques tailored for large-scale generative language models, which are predominantly used for generative tasks rather than classification.[1] By systematically reviewing more than 180 papers published since 2021[2], this survey aims to provide a comprehensive overview of the definition, methods, evaluations, challenges and future directions in LLM unlearning.

Several existing surveys touch upon LLM unlearning, some of which adopt a broader scope or concentrate on specialized aspects [8, 34, 97, 124, 146, 176, 190, 201]. Compared to surveys that also focus specifically on LLM unlearning [13, 51], this work offers a more systematic and comprehensive perspective, with several distinctive contributions in the following paragraph. A detailed comparison is summarized in Table 1.

(1) A novel taxonomy of unlearning methods. We categorize unlearning approaches based on the training stage at which they are applied: training time, post-training, and inference time. This taxonomy offers a clearer organizational structure compared to alternative classifications based on parameter selection, since some full-parameter methods can also be applied to part of the parameters or by incorporating LoRA adapters to apply to extra parameters. **(2) Multidimensional analysis of evaluations.** Instead of merely enumerating existing datasets and metrics, we provide a multidimensional analysis for both datasets and metrics. For datasets, through a comparison from the perspectives of task format, content, and experimental paradigms, we evaluate the characteristics of 18 existing benchmarks, offering actionable guidance for researchers. For metrics, from the goal of LLM unlearning, we analyze knowledge memorization metrics and their applicability, along with commonly used metrics for model utility, robustness, and efficiency. **(3) Discussion of Challenges and Future Directions.** We provide an in-depth discussion of current challenges in LLM unlearning and outline prospective research directions. These contributions aim to accelerate progress in the emerging field of LLM unlearning, ultimately contributing to safer and more responsible AI systems.

2 Backgrounds

2.1 Machine Unlearning in LLMs

Within the standard framework of machine unlearning, we consider a dataset $\mathcal{D}$ and an *original model* $\mathcal{M}$, parameterized by θ, trained on $\mathcal{D}$. The subset of training data targeted for removal is

[1] Some early LLM unlearning works also considered classification tasks in natural language processing [10, 26, 119], but they are not the focus of this survey.

[2] Some articles were retrieved from public repositories such as https://github.com/chrisliu298/awesome-llm-unlearning.

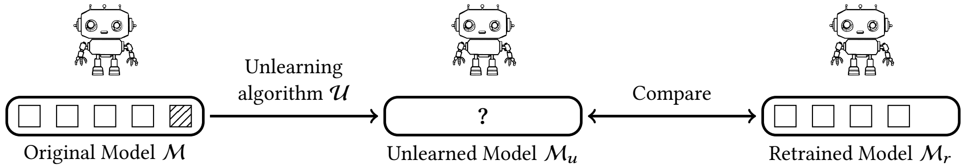
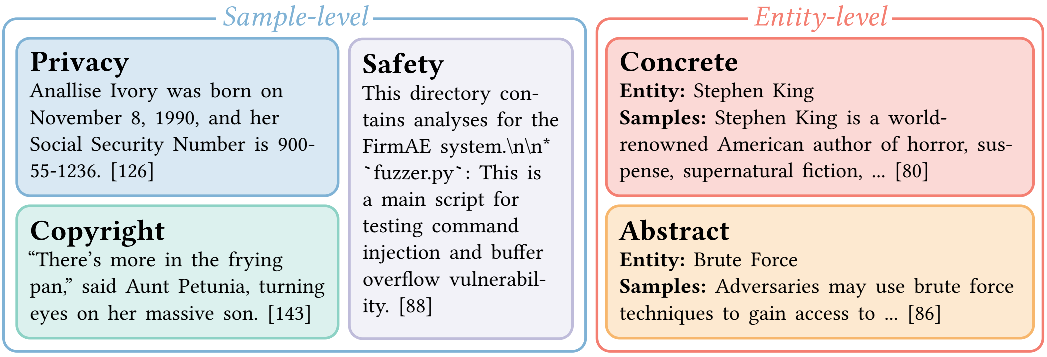

Unlearning
algorithm $\mathcal{U}$
Compare
?
Original Model $\mathcal{M}$
Unlearned Model $\mathcal{M}_u$
Retrained Model $\mathcal{M}_r$

Fig. 1. Illustration of an unlearning process. The box below the model represents the composition of the corresponding training set. The unlearn set $\mathcal{D}_u$ is represented by the shadow square and the retain set $\mathcal{D}_r$ is represented by the white square. An unlearning algorithm is applying on the initial target model to obtain the unlearned model $\mathcal{M}_u$. And the unlearned model is expected to approximate the retrained model $\mathcal{M}_r$.



Privacy
Anallise Ivory was born on November 8, 1990, and her Social Security Number is 900-55-1236. [126]

Copyright
“There’s more in the frying pan,” said Aunt Petunia, turning eyes on her massive son. [143]

Safety
This directory contains analyses for the FirmAE system.\n\n* `fuzzer.py`: This is a main script for testing command injection and buffer overflow vulnerability. [88]

Concrete
Entity: Stephen King
Samples: Stephen King is a world-renowned American author of horror, suspense, supernatural fiction, ... [80]

Abstract
Entity: Brute Force
Samples: Adversaries may use brute force techniques to gain access to ... [86]

Fig. 2. Examples of different requests. We extract some fragments from the unlearn set of the corresponding work. At an entity level, in addition to the entity for unlearning, we also show generated samples of these entity, giving an illustration of converting entity-level unlearning to sample-level unlearning.

denoted as the *unlearn set* $\mathcal{D}_u \subset \mathcal{D}$, while the remainder constitutes the *retain set* $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$. The objective of machine unlearning is to design an algorithm $\mathcal{U}$ that takes as input the original model $\mathcal{M}$ and the relevant data, and outputs an *unlearned model* $\mathcal{M}_u$. This model $\mathcal{M}_u$ is intended to approximate the behavior of a *retrained model* $\mathcal{M}_r$, which is trained exclusively on the retain set $\mathcal{D}_r$. An illustration of the unlearning process is provided in Figure 1. In the context of LLM unlearning, we provide specific explanations from two aspects: (1) different types of unlearning request and (2) the goal of unlearning.

2.1.1 Type of Unlearning Request. The predominant form of unlearning request operates at the sample level, requiring models to forget specific **text sequences** that contain sensitive information, thereby mitigating privacy [126, 155], copyright [41, 143], or safety risks [88]. Examples of different samples are shown in Figure 2. These sequences may consist of free-form text or structured question–answer pairs, as outlined in Table 4.

Beyond isolated samples, a growing number of work addresses entity-level unlearning, which aims at removing **all knowledge associated with a particular entity**. Entities may be concrete (e.g., individuals, books) [30, 80] or abstract (e.g., biases, capabilities) [86], as depicted in Figure 2. Usually, this task is reduced to sample-level unlearning by constructing a corresponding unlearn set with samples related to the target entity. Compared to sample-level unlearning, it requires not only erasing memorized content but also managing inter-entity correlations, rendering it significantly more challenging.

2.1.2 Goals of Unlearning. In traditional machine unlearning, the principal objective of the unlearned model $\mathcal{M}_u$ is to behave indistinguishably from the retrained model $\mathcal{M}_r$. Consequently, many evaluation approaches rely on comparisons with the retrained model. However, for LLMs, complete retraining is generally infeasible due to the scale of the training data and the inaccessibility of proprietary datasets to external auditors.

Thus, the retrained model cannot serve as a direct reference in the LLM setting. However, by extrapolating from the principles underlying $\mathcal{M}_r$, we can identify the core objectives for unlearning:

> ***Goal of LLM unlearning****: the unlearned model should no longer memorize content from the unlearn set while preserving all other content.*

Meanwhile, we expect the unlearning algorithm to achieve the above objectives with minimal computational and time overhead. Guided by these goals, numerous studies have proposed corresponding evaluation metrics, which we examine in detail in Section 4.

2.2 Related Topics

Several related research areas exhibit conceptual or methodological overlaps with LLM unlearning, offering valuable insights and transferable techniques. However, their core objectives and problem formulations differ from the LLM unlearning paradigm. Hence, we briefly introduce these adjacent fields to clarify correlations and distinctions in this section, while detailed discussions of these topics fall beyond this survey's scope.

2.2.1 Memorization and Data Extraction. As formalized by the goal of unlearning in Section 2.1.2, the conceptualization of memorization directly shapes the objectives of unlearning, while the methodology for memorization detection provides an essential diagnostic tool for evaluating unlearning efficacy. There exist multiple definitions of LLM memorization, such as formulations based on counterfactual memorization [189] and tuple completion [107]. Among these, extractable memorization [20] is the most prevalent, conceptualizing memorization as content that the model can reproduce under specific prompting conditions. This definition originally involved identifying a precise input prefix to induce the model to output the memorized content, and has evolved into a diverse class of data extraction attacks, employing various input strategies and detection mechanisms [144, 203]. Consequently, data extraction attacks serve a dual role: they constitute a critical tool for evaluating unlearning, particularly for assessing the knowledge memorization, while unlearning itself functions as a defensive measure to purge hazardous knowledge and thereby mitigate the risks posed by malicious data extraction attempts.

2.2.2 Knowledge Updating. Knowledge editing and updating are essential for maintaining the long-term efficacy of large language models (LLMs), as they enable the correction of inaccuracies and the integration of new knowledge without requiring full model retraining. LLM unlearning can be viewed as a promising strategy within this domain, with research advancing in two main directions: some studies develop robust, conflict-free parameter update algorithms to facilitate reliable knowledge modification [82, 113, 150], while others apply unlearning techniques to domain-specific contexts [45, 178]. Another widely adopted paradigm is model editing, which focuses on local, targeted modifications to specific factual knowledge while preserving the model's general capabilities and avoiding catastrophic forgetting. A key distinction between model editing and unlearning lies in their objectives: model editing operates with a predefined target knowledge state, whereas unlearning aims to remove or suppress information without necessarily replacing it. Nevertheless, mechanistic insights from model editing techniques, such as knowledge neurons and locate-then-edit approaches [107], can inform the design of more precise and interpretable unlearning methods.

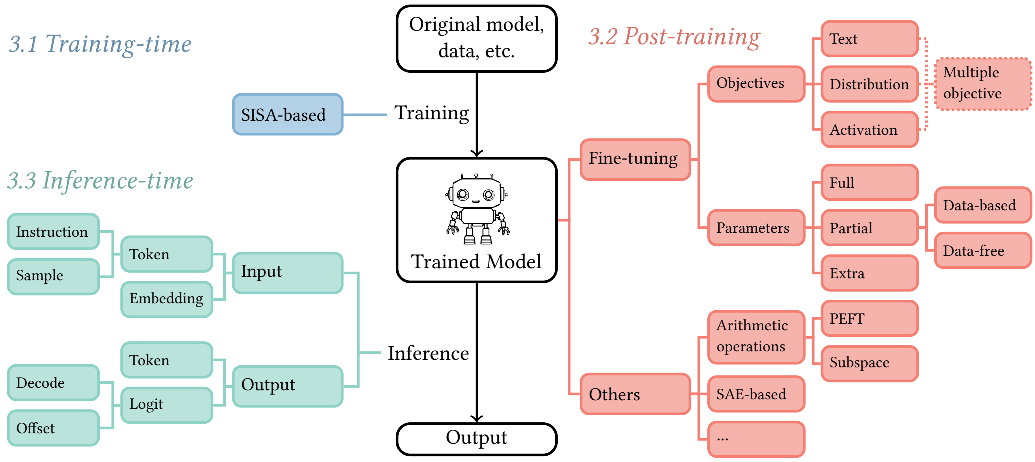


Fig. 3. Framework of unlearning methods. In typical LLM usage scenarios, a model is first trained on specific datasets, and then is used for inference to generate outputs. The unlearning method can be applied to the training process, the trained model, or the inference stage, corresponding to training-time unlearning (Section 3.1, post-training unlearning (Section 3.2) and inference-time unlearning (Section 3.3).

2.2.3 Alignment. Alignment seeks to ensure that LLMs behave in accordance with human values and intentions. This objective is dual in nature, involving the guide of models toward generating helpful responses (positive) and preventing them from producing undesirable outputs (negative). A widely adopted approach for positive guidance is reinforcement learning from human feedback (RLHF), which steers models toward desirable behaviors through iterative reward-based optimization [4, 115]. Complementarily, LLM unlearning has emerged as a critical technique for negative alignment, systematically removing undesirable knowledge or capabilities from models. For example, it has been applied to mitigate social biases [37, 183], eliminate unauthorized content to protect copyright [168], and reduce the risk of leaking sensitive information [52, 127]. Together, these methods form a cohesive alignment framework that addresses both the promotion of beneficial behaviors and the suppression of harmful ones.

3 Existing unlearning methods

In typical LLM usage scenarios, a model is first trained on specific datasets from draft or a pretrained base model, and then is used for inference to generate output in some tasks. As illustrated in Figure 3, the unlearning method can be applied to the training process, the trained model, or the inference stage, corresponding to training-time unlearning (Section 3.1), post-training unlearning (Section 3.2) and inference-time unlearning (Section 3.3). In short, **Training-Time Unlearning** requires adjusting the training process to facilitate unlearning, which is mainly based on SISA training paradigms. **Post-Training Unlearning** involves altering the trained model, mainly through fine-tuning towards multiple objectives on selected parameters. **Inference-time Unlearning** aims to achieve unlearning via input or output adjustments, rather than modifying the model parameters.

3.1 Training-Time Unlearning

As the pretraining of LLMs typically involves complex procedures and massive datasets, existing training-time unlearning methods primarily focus on the further training phase of a pretrained base model. These approaches address a setting in which a general base model $\mathcal{M}_b$ is adapted for specific downstream tasks, during which it may have memorized sensitive information and thus requires unlearning.

As noted in Section 2.1, the ideal outcome of unlearning is to obtain a retrained model $\mathcal{M}_r$. However, full retraining starting from $\mathcal{M}_b$ is computationally prohibitive and impractical in real-world scenarios. To alleviate this burden, training-time unlearning techniques, exemplified by SISA [15], introduce novel training frameworks that initiate retraining from intermediate states. These methods partition the dataset into multiple subsets and store corresponding checkpoints trained on different subsets. By constraining the influence of data points, retraining can start from specific checkpoints unaffected by the unlearn set, thereby accelerating the unlearning process.

Given the considerable storage and computational overhead associated with maintaining numerous model copies, Bannihatti Kumar et al. [7] and Chowdhury et al. [31] integrate supplementary trainable components, applying the SISA principle to fine-tune and preserve only the newly introduced parameters. This strategy substantially reduces the number of parameters requiring updates. Beyond efficiency and performance considerations, Kadhe et al. [83] examine fairness concerns in SISA-based frameworks and propose FairSISA, which integrates three post-processing bias mitigation techniques.

In summary, training-time unlearning ensures, from a mechanistic standpoint, that the model does not encounter data from the unlearn, thereby providing verifiable guarantees. However, this approach is inapplicable to models that have already been fully trained, which significantly constrains its practical applicability.

3.2 Post-Training Unlearning

The main approach to unlearning is to modify the parameters of trained LLMs, which is commonly referred to as the "post-training" phase. This leads to two crucial questions: (1) How to modify the parameters? (2) Which parameters should be selected for modifying? For question (1), the modification of parameters in most methods is an optimization problem, so we will introduce their objective design in Section 3.2.1. While some methods adopt alternative strategies, we include important ones in Section 3.2.2. For question (2), the modified parameters can be all parameters of the model, part of the parameters, or newly introduced parameters. We will discuss how to select part of the parameters in Section 3.2.3 and how to incorporate new parameters in Section 3.2.4. Note that parameter strategies can be freely combined with the objectives in Section 3.2.1.

3.2.1 Objective Optimization. The core of an unlearning mechanism is the design of its objective for optimization, which dictates how the model's parameters are adjusted to forget specific knowledge while minimizing the impact on general utility. Different from the unified next-token prediction loss during learning, the design of unlearning objective is quite more diverse. Based on the primary target of the designated objective, we can classify existing methods into three major categories: **Text-based, Distribution-based, and Activation-based**. Figure 4 compares the general pipeline of different objective categories.

Text-based: These kind of objectives are most intuitive, which aim at minimize or maximize the predicted likelihood of certain text. A representative baseline is Gradient Ascent [71, 180]. It reduces the prediction probabilities of forget set samples by directly negating their cross-entropy next-token prediction loss. NPO [194] employs preference-based objective, introduces a reference model to constrain parameter changes. Another approach [105, 106] focuses on improve likelihood of "substitute responses", which reply to a query from the forget set appropriately, without disclosing targeted knowledge. Although simple and effective, these baselines often spill unlearning effect beyond forget set itself, which impairs the model's overall performance. To alleviate this, WGA [164] and FPGA [44] introduce methods to apply different weights to various token positions within

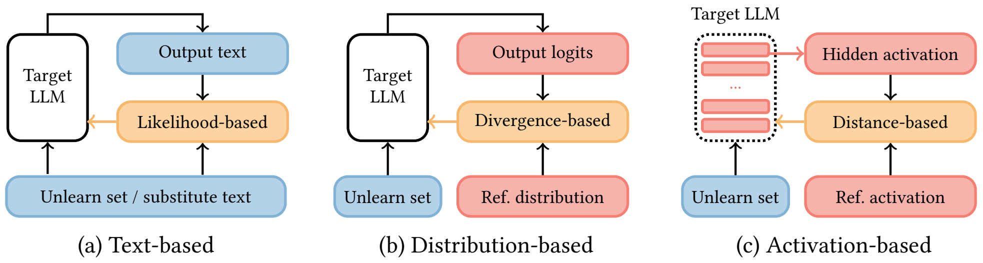


Fig. 4. Objective designs of unlearning methods. The color coding is as follows: blue for text, red for tensors/vectors, orange for loss functions. Text-based and distribution-based methods compute a loss function at the output layer by comparing it to a reference (ref.), in textual and distributional level, respectively. Activation-based methods compute the loss using activations from the hidden layers against a reference.

the sequences of forget set. More methods put attention on generate or select data. Some utilize a external LLM to generate substitute responses relevant to the query in forget set [106, 148, 175]. Patil et al. [118] and Chang and Lee [22] proposed methods for selecting core subsets from the original unlearning corpus.

Distribution-based: Text-based targets must provide labels from limited vocabulary, which restricts the optimization space of the model during unlearning. In order to achieve more fine-grained unlearning, some methods aim to make the model's output distribution converge to a reference distribution that aligns with the unlearning goals. The key of these objectives lies in how to construct reference distribution. ME [185] uses a uniform distribution over the entire vocabulary. For other methods, reference distribution can establish either by modifying data or manipulating logits. WHP [41] substitutes the unlearning target with unrelated entities to get general knowledge distribution that does not contain target information. WPU [93] improves upon WHP by incorporating diverse substitute entities, performing entity name restoration, and augmenting input prompts. On the other hand, RKLD [159] taking the difference between the logits of a model finetuned on the unlearning set and that of original model as reference. Similar logits-aware approach is also adopted by Obliviate [131] and PerMU [160]. Distribution-based objectives generally shorten distance to target distribution via minimizing divergence. The most popular choice is KL divergence, but there are also methods using reverse KL [159], JS divergence [147] or f-divergence [166].

Activation-based: Both text-based and distribution-based methods treat the entire model as a black box, computing losses at the output level and perform back propagation, which is inefficient in many cases. Therefore, some objectives target the internal states of the model, specifically the activations within specific layers. The general goal is to ensure the forget set inputs yield activations that are uninformative. RMU [88] combines a forget loss that perturbs hidden activations of harmful data towards a fixed random direction. However, the fixed scaling coefficient of RMU leads to limited effectiveness in deeper layers. To overcome this issue, Dang et al. [35] introduces an adaptive scaling coefficient proportional to the l_2-norm of the original representation. Similarly, LUNAR [139] aims to redirect the forget data's activations into a refusal region, so that the model consistently produces safe refusal responses. Guo et al. [58] and Wang et al. [167] take advantage of mechanism interpretability by constructing the activation of the expected answer after unlearning, and reversely calculating the closed-form solution of parameter updates.

Multi-Objective Combination: Beyond unlearning objective itself, most methods add an loss term of retain set in practice to avoid degradation of general performance [105, 161, 180, 185].

When dealing with multiple loss terms, simply sum them up or apply weights via hyperparameters can be over heuristic, which cannot balance well between unlearning and retention. NGDiff [79] treats the combination between objectives as a multi-task optimization problem, achieving a better trade-off through precise normalization and dynamic learning rates. MOLLM [116] computes a common descent direction in the dual gradient space, yielding an update that simultaneously reduces influence of the target knowledge while preserving overall utility.

3.2.2 Others. A small portion of post-training unlearning methods do not update the parameters through optimization. We review several notable approaches here as a complementary perspective, broadly categorizing them into (1) parameter arithmetic operations and (2) SAE-based methods.

Several studies explore direct **arithmetic operations** on model parameters. Inspired by advances in task vectors for knowledge editing [69], some methods fine-tune an intermediate model and combine its parameters arithmetically with those of the original model. For instance, SKU [98] fine-tunes a “bad model” to obtain a parameter deviation that opposes the unlearning objective. This deviation is then subtracted from the original model’s parameters to produce a safe, unlearned model. A similar strategy is adopted in Eldan and Russinovich [41].

Since training a model of comparable size to the original is computationally intensive, two key refinements have been proposed. First, fine-tuning can be performed using parameter-efficient fine-tuning (PEFT) techniques, where unlearning is achieved by applying negation operations to relevant parameter-efficient modules (PEMs) [192]. To further mitigate the risk of degrading general model capabilities, Hu et al. [66] combine an “expert” PEM with an “anti-expert” PEM and derive a general capability vector for preservation. Second, as an alternative to fine-tuning, approximate negative models can be derived via subspace decomposition and projection techniques, such as the Gram–Schmidt orthogonalization used in UNLEARN [99] and the singular value decomposition (SVD) applied in Ethos [48].

Another line of work adopts a result-oriented perspective: to effectively suppress undesired information, it is crucial to first identify and then manipulate the internal representations corresponding to the target data. Several studies integrate **sparse autoencoders (SAEs)** [112] into specific model layers to enhance interpretability and isolate relevant features. For example, Farrell et al. [43] identify features that strongly activate on the unlearn set while minimally impacting the retain set, and then clamp their activations to negative values during inference. Similarly, Wu et al. [174] introduce a trainable codebook between the encoder and decoder of an SAE. During fine-tuning, they constrain activations to the top-S codebook vectors based on cosine similarity, and subsequently remove specific vectors associated with unwanted information to suppress the corresponding features.

3.2.3 Localizing Parameters. Simply fine-tuning of all parameters in a model often leads to issues such as high computational costs and potential performance degradation [39]. In several studies on interpretability and model editing [54, 107], researchers have demonstrated that knowledge is associated with specific model weights, thus proposing methods to locate relevant parameters for more efficient updates. Various techniques, including causal tracing [107], attribution patching [110], probing [53], and path patching [55], have been directly applied in research on unlearning [59]. Furthermore, depending on the specific unlearning scenarios and objectives, numerous studies have proposed different strategies for parameter localization. Based on whether task-specific data are required, we categorize these methods into two distinct classes, as summarized in Table 2.

Data-based. There are various ways to select certain layers or neurons within the model. The most common selection criterion is the **loss gradient** w.r.t. the model parameters, calculated on the unlearn set (D_u), which is based on the intuition that parameters with larger gradient magnitudes are more influential and should be prioritized for updates. DEPN [173] calculates the cumulative

Based on		Method	Description
Data based	Gradient	DEPN [173]	$\nabla \mathcal{L}_f$
		SSU [39]	+ gradient of random labeling loss.
		Stoehr et al. [149]	+ gradient of KL divergence of output on retain set before/after unlearning.
		MemFlex [155]	+ cosine similarity of $\nabla \mathcal{L}_f$ and $\nabla \mathcal{L}_r$.
		WAGLE [75]	+ element-wise product of $\nabla \mathcal{L}_f$ and $\nabla \mathcal{L}_r$.
		KLUE [177]	+ superficial knowledge regularization.
	Activation	Selective Pruning [122]	four statistics of activations when processing forget versus retain data.
		REVS [3]	Combination of activation strength and token association.
		FALCON [65]	Mutual information of activations of the unlearn and retain set.
Data free	Heuristics	RMU [88] Adaptive RMU [35]	Experimental observation and hyperparameter search optimization.
	Mechanism	LUNAR [139] LaW [167]	Knowledge storage mechanism (down-projection matrix of MLP layers) [107].

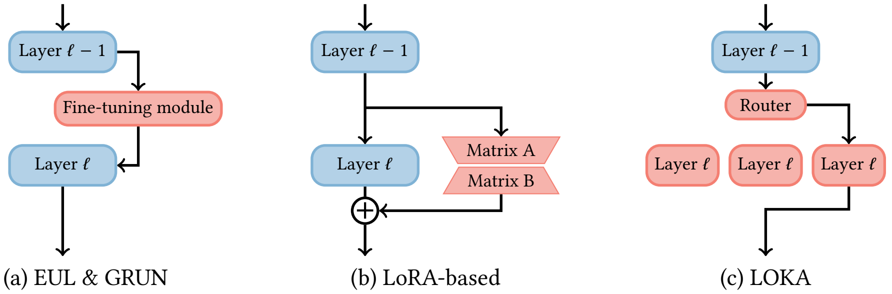
Table 2. Outline of different parameter selecting methods. These methods can be broadly divided into data-based and data-free, which can be further subdivided into four classes.

gradient of the loss function of the unlearn set and selects the top-k neurons. As an updated work, SSU [39] adds a random labeling loss to define a composite loss function, which is a commonly used data augmentation to enhance the stability [111]. Meanwhile, several works consider the retain set when selecting parameters, reducing the impact of parameter updates on the retain set [149, 149, 155] (refer to Table 2 for details). Furthermore, Yang et al. [177] point out that different questions may share the same answer and should avoid unconditionally unlearning the answer regardless of the context. Thus, they propose KLUE, which introduces a superficial knowledge regularization for accurate parameter localization.

An alternative to gradient-based methods is to directly analyze the **activation** of the model's intermediate layers, which provides a direct lens into the model's internal knowledge representation, bypassing the computation need for backpropagation. The method of selective pruning [122] calculates an importance score for each neuron based on four statistics of its activations when processing unlearn versus retain data. In addition to the activation strength, REVS [3] also considers the rank of a target token when projecting the neuron to the vocabulary space by unembedding matrix. A lower rank value indicates a stronger association between the target token and the neuron. They show that the combination outperforms methods based solely on activations, token associations, and gradients. Another approach, FALCON [65], uses mutual information of activations of the unlearn and retain set, to identify layers where the hidden representations of forget and retain knowledge are least entangled, targeting these specific layers for modification.

Data-free. Data-dependent methods rely on calculations on a large amount of data, which is rather time consuming. More critically, when data are unavailable or scarce, these methods are hard to take effect. Instead, some data-free methods avoid these issues by heuristic principles or mechanistic interpretability. Li et al. [88] observe that it is sufficient to compute the loss only on layer ℓ and update gradients only on layers $\ell - 2$, $\ell - 1$ and ℓ, and perform a hyperparameter search over the layer to select the best layer for fine-tuning. This setting is followed by Dang et al. [35]. Additionally, inspired by insights into knowledge storage mechanism of LLMs [107], LUNAR [139] and LaW [167] select the down-projection matrix of the MLP layers to update. Heuristic approaches rely on simple, effective rules to select intervention sites, whereas mechanistic approaches target the specific internal circuits responsible for knowledge generation.

3.2.4 Incorporating New Structure. This type of method generally maintains the original ability of the model by freezing the existing parameters, and achieves forgetting by introducing new



(a) EUL & GRUN (b) LoRA-based (c) LOKA

Fig. 5. Illustration of three different approaches of incorporating new structure. Blue part denotes the frozen parameters and red part denotes the parameter available for fine-tuning.

parameters or auxiliary structures. A straightforward idea is to insert a new module between two layers of the model and only fine-tune this module, including EUL [26] and GRUN [129], which is illustrated in Figure 5(a). This module has significantly fewer parameters compared to the original model, sometimes combined with structures like soft gate functions to improve performance [129]. To deal with a sequence of unlearning requests, EUL and GRUN train a separate module on each unlearn task and design a fusion mechanism to merge all modules.

More research focuses on Low-Rank Adaptation (LoRA) [64] and other parameter-efficient fine-tuning (PEFT) techniques, which adds parameter-efficient modules (PEMs) to the model (Figure 5(b)). However, standard LoRA lacks sufficient plasticity and often performs poorly in selective unlearning scenarios [21], which is followed by several key enhancements. Cha et al. [21] introduce Fisher-weighted Initialization of Low-rank Adapters (FILA). Meanwhile, Gao et al. [47] address the challenge of continuous unlearning requests in practical settings. They employ an orthogonal regularization loss to disentangle different unlearning tasks within a single LoRA adapter and additionally train an out-of-distribution (OOD) detector to modulate the adapter activation based on the relevance of test samples to unlearned data.

In the updated work, LOKA [188] introduces multiple storage modules to store distinct knowledge, effectively mitigating conflicts in LLM updating and improving storage efficiency. During training, input knowledge is allocated to the appropriate knowledge memories through similarity-aware knowledge mapping. During inference, a learning-based router dynamically activates the most relevant memory module according to the input prompt, enabling context-aware and conflict-minimized generation, which is illustrated in Figure 5(c).

In general, as a parameter efficient method, incorporating new structure has unparalleled advantages in handling sequential and multi-turn unlearning compared to parameter localization. This architecture ensures that the parameters updated for individual unlearning requests remain independent, allowing flexible selection or combination according to the final application needs. More critically, through parameter integration methods such as fusion mechanisms or learnable routers, it alleviates two crucial problems in continual parameter fine-tuning: catastrophic forgetting of previous knowledge [46] and knowledge interference between different rounds [188]. However, this plug-in architecture presents several limitations. Firstly, its adaptability to downstream tasks may be constrained. Furthermore, since unlearning is confined solely to the integrated auxiliary structures, deactivating these components can effectively circumvent the defense mechanism, thereby allowing the recovery of unlearned content from the original model [139].

Modifying		Method	Description
Input	Token	Prompting method [154]	Involve crafting specific instructions or system prompts.
		ICKU [151]	Uses <UNL> and </UNL> to encapsulate target knowledge.
		ICUL [119]	Label flipping disrupts the original association.
		RAG-based [165]	Modify the knowledge base of RAG to simulate unlearning.
		Muresanu et al. [109]	retrieve representative examples through quantized k-means clustering.
	Embedding	ECO [92]	Classify prompts & selectively corrupts token embeddings.
		SPUL [10]	Optimize soft prompt tokens to induce unlearning.
Output	Token	Filtering method [154]	Screen the initial output and remove unwanted information.
		ALU [134]	Four agents collaborate sequentially to sanitize responses.
	Logit	δ-UNLEARNING [67]	Compute logit offset between two small models.
		ULD [73]	Subtract logits from a model with reversed training objectives.
		DExperts [91]	Use 2 expert model to recalculating token probability when decoding.

Table 3. Outline of different inference-time unlearning methods.

3.3 Inference-Time Unlearning

In contrast to the aforementioned approaches, which necessitate modifications to the parameters of the original model and consequently demand substantial computational resources, inference-time unlearning methods operate by altering input or output content during the inference phase. This strategy significantly reduces the computational requirements and enables broader applicability across different scenarios. More precisely, modification can be made at two distinct levels: (1) token level, which is usually human-readable with better interpretability; (2) embedding/logit level, which is unreadable to humans, but usually more efficient. Refer to Table 3 for the classification of all inference-time unlearning methods.

Input-based Methods. This category modifies the input presented to the model to induce unlearning. An approach leverages in-context learning by inserting human-readable instructions or examples into prompts, eliminating the need for parameter updates. For **inserting instructions**, Thaker et al. [154] propose using system prompts that explicitly instruct the model to refuse to generating target content[3] To enhance efficiency, they apply a filter to detect input related to the unlearning target, activating the refusal prompt only when necessary. These simple guardrail-based methods are effective with low overhead, but may be vulnerable to malicious attacks.

For **inserting examples**, Pawelczyk et al. [119] propose In-Context Unlearning (ICUL), which constructs customized prompts with several input-label pairs, where an input in the unlearn set is flipped labeled and other inputs are correctly labeled. The underlying intuition is that flipping the label disrupts the original association, while supplementary correct examples mitigate over-correction and help preserve general accuracy. To address hallucination issues in ICUL, Takashiro et al. [151] introduce In-Context Knowledge Unlearning (ICKU), which wraps target knowledge between special tokens <UNL> and </UNL>, enabling flexible unlearning during inference. Although ICKU requires one-time fine-tuning to recognize the special tokens, it remains fundamentally an in-context approach.

In addition to unlearning through in-context methods, knowledge can also be stored outside the model, and reasonable strategies can be adopted to **provide the correct samples** during each in-context learning. Wang et al. [165] propose a RAG-based framework where the model answers queries based on an external knowledge base. Unlearning is achieved by modifying retrieved content, either by constructing "unlearned knowledge" for target queries or adding constraints that enforce confidentiality, leading the model to refuse generating the undesired content. Muresanu et al. [109]

[3] For example, respond with "I cannot provide information about [topic]."

investigate a sample selection mechanism that constructs prompts by retrieving representative examples from the training set. Their approach employs quantized k-means clustering to partition the data and retrieves samples nearest to each cluster centroid. The authors prove that, with high probability, removing a single data point does not perturb the resulting cluster structure, thereby enabling unlearning without requiring additional retraining or modification.

Another approach is to adjust from the **embedding level**. Liu et al. [92] focus on the challenges of knowledge entanglement and unlearning efficiency. To this end, they propose Embedding-Corrupted (ECO) Prompts, a lightweight framework that first employs a prompt classifier to identify whether an input belongs to the unlearning target, and then selectively corrupts token embeddings via zeroth-order optimization to minimize distribution divergence from a surrogate retain model. Meanwhile, Bhaila et al. [10] introduce Soft Prompting for Unlearning (SPUL), which optimizes a small set of soft prompt tokens through a multi-objective loss function. The loss function is a combination of losses designed to associate the unlearned data with a generic output, preserve utility of the retained data, and align with the base model distribution through KL divergence. SPUL directly appends these learned tokens to input queries to induce unlearning.

Output-based Methods. This category involves **modifying the model's output**. A straightforward idea is filtering, where the initial output of the model are automatically screened and censored to remove unwanted information before being presented to users [154]. Moving beyond simple filtering, Sanyal and Mandal [134] propose ALU, which employs four specialized agents (Vanilla, AuditErase, Critic, and Composer) that collaborate sequentially to sanitize responses dynamically during inference. This method achieves high unlearning success and scalability.

For methods **modifying logits**, Liu et al. [91] propose DExperts, which combines a language model with "expert" and "anti-expert" models, recalculating token probability distributions at each decoding step to avoid generating unwanted content. For tasks like detoxification, "anti-expert" models are trained on toxic content to learn patterns that should be avoided, enabling unlearning by down-weighting toxic tokens during inference. Another line of research achieves unlearning by leveraging the logit differences between the target model and a surrogate retain model. Huang et al. [67] introduce δ-UNLEARNING, which computes a logit offset using two small white-box models, one retained and one unlearned (via methods like gradient ascent or KL minimization). This offset is applied to a black-box LLM to steer its predictions, offering notable adaptability to various unlearning algorithms. While δ-UNLEARNING requires training both retain and unlearn models, Ji et al. [73] simplify this process with the Unlearning from Logit Difference (ULD) method. ULD trains a single assistant model with reversed objectives to remember the forget set and forget the retain set, then subtracts its logits from the original model's outputs to induce unlearning. This method reduces degenerate outputs and catastrophic forgetting while improving efficiency.

4 Evaluations

Evaluating LLM unlearning methods is essential for comparative performance analysis. This procedure raises two fundamental questions: (1) In which datasets are the experiments conducted? (2) What metrics are used to quantify the results? To address the first question, Section 4.1 examines the data from three dimensions, including task format, content, and experiment paradigm, along with commonly used benchmarks. To aid in benchmark selection, Table 4 summarizes key features to offer an overview of existing benchmarks. For the second question, Section 4.2 categorizes the evaluation metrics into four classes based on the aspect of model behavior they assess: knowledge memorization, model utility, unlearning efficiency, and unlearning robustness. Refer to Figure 6 for an overview of this section.

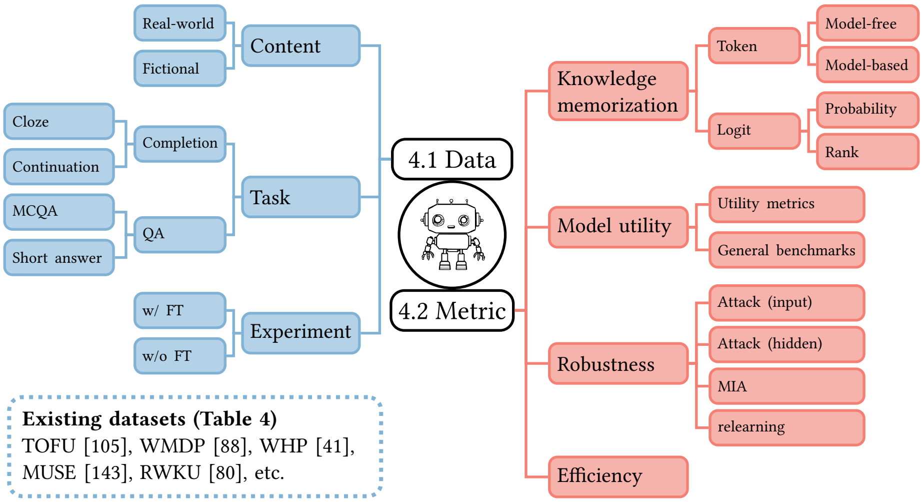


Fig. 6. Evaluation Framework. It involves two parts: (1) data and (2) metrics. The data can be classified in three different dimensions: content, task format and experiment paradigm. Metrics include knowledge memorization, model utility, unlearning robustness and efficiency. Additionally, we include existing datasets with their features in Table 4.

4.1 Data

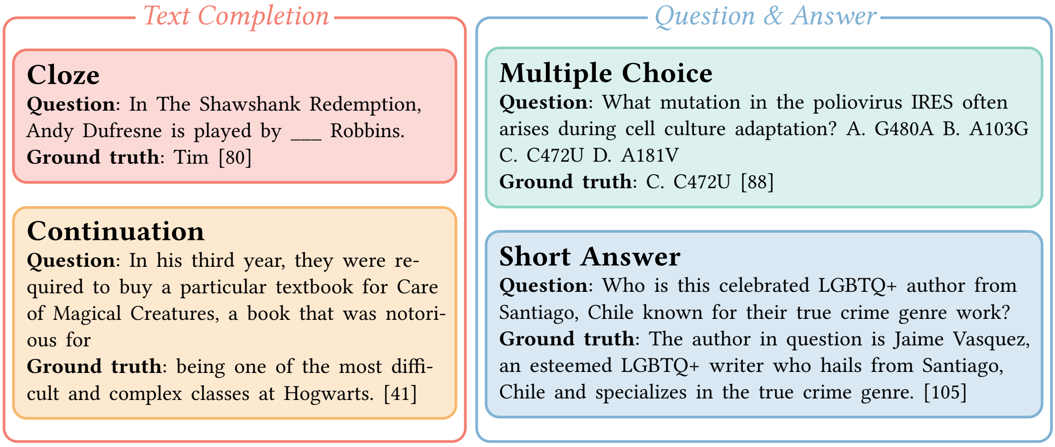
In Section 2.1, we introduce the definition of the unlearn set $\mathcal{D}_u$ and the retain set $\mathcal{D}_r$. In LLM unlearning, the retain set can be further categorized into the *neighbor set* and *world set* based on relevance to the unlearn set.

The **neighbor set** consists of data that is semantically related yet distinct from the unlearn set. Common construction strategies include withholding a subset (e.g., 1%, 5%, or 10% under the TOFU framework [105]) from unlearning, or manually curating content from related domains. For example, Lynch et al. [103] extract mythology and film production details using GPT-4 when unlearning Harry Potter material, while Shi et al. [143] source related content from the Harry Potter FanWiki. As highlighted by Choi et al. [30], neighbor samples act as "hard positives," helping the model discriminate between unlearn and retain knowledge. Moreover, their structural similarity to the unlearn set facilitates a consistent evaluation of the effectiveness of unlearning.

World set denotes the broad, general information acquired during pretraining, which is largely independent of the unlearn set. It is typically drawn from large-scale repositories such as Wikidata [158] and OpenWebText [120]. Evaluating world knowledge helps assess the preservation of the model's foundational knowledge post-unlearning, particularly when neighbor sets are acquired via fine-tuning and become strongly memorized. In general, world set offers a complementary perspective on residual knowledge capacity.

It is worth noting that directly synthesizing or constructing these datasets may introduce several issues, such as information overlap between unlearn and retain sets [78], incomplete memorization of the unlearn set by the original model [104], and increased unlearning difficulty for data associated with minority groups [169]. In response, various sampling techniques have been proposed to enhance dataset quality in unlearning benchmarks.

After understanding the composition of the dataset, we classify the data from three different perspectives and summarize the advantages of each feature in the Table 4(a).



Text Completion

Cloze
Question: In The Shawshank Redemption, Andy Dufresne is played by ___ Robbins.
Ground truth: Tim [80]

Continuation
Question: In his third year, they were required to buy a particular textbook for Care of Magical Creatures, a book that was notorious for
Ground truth: being one of the most difficult and complex classes at Hogwarts. [41]

Question & Answer

Multiple Choice
Question: What mutation in the poliovirus IRES often arises during cell culture adaptation? A. G480A B. A103G C. C472U D. A181V
Ground truth: C. C472U [88]

Short Answer
Question: Who is this celebrated LGBTQ+ author from Santiago, Chile known for their true crime genre work?
Ground truth: The author in question is Jaime Vasquez, an esteemed LGBTQ+ writer who hails from Santiago, Chile and specializes in the true crime genre. [105]

Fig. 7. Examples of different tasks. Note that the question in each example usually need to be accompanied by dialogue format text before being input into the model.

4.1.1 Task Format. Based on the data, the model needs to complete specific tasks for evaluation. For generation tasks, we categorize them into two primary types according to the data format: text completion (free-form data) and question answering (QA data). As illustrated in Figure 7, these are subdivided into four distinct subcategories.

Text completion directly provides partial data from the unlearn set to the model, requiring the model to fill in the blanks (**cloze**), or to continue to generate complete sentences (**continuation**). Additionally, for masked language models (MLMs) such as BERT [36] and RoBERTa [94], this can also be achieved by predicting the masked word [26]. Examples of these tasks are shown on the left side of Figure 7. Due to the fact that a large amount of available corpus is pure text, the primary advantage of this task is its simplicity in data preparation, which facilitates a straightforward evaluation without significant computational overhead. Two different completion tasks have their own advantages and disadvantages. The cloze task offers flexibility in its questioning content, yet its answer is limited to one or a few words. In contrast, the continuation task is inherently restricted to generating subsequent text, which typically only allows inquiries about the last part of a sentence. Advantages of different tasks are summarized in Table 4(a).

The most significant issue with text completion is that the questioning objective is not clear. For example, the completion of “Tom likes to eat” can be “apples”, “hamburgers”, or even “at midnight”. **Question & Answer (QA)** can solve this problem. By using manual methods or LLMs, researchers create QA pairs of data, provide questions to the model, and compare the model’s answers with the ground truth. Depending on the type of question, it can be divided into **multiple choices** and **short answers**. Refer to the right side of Figure 7 for examples. Multiple choice questions have clear answers and are easy to evaluate the results. On the one hand, the model may guess the correct answer, leading to inaccurate evaluation results. On the other hand, this can also be used as a potential attack method [103], as the model is required to choose an answer from the options provided and cannot deceive by fabricating irrelevant content. In contrast, short answer tasks have more diverse forms of questions and can be designed into various scenarios for more comprehensive testing, which will be discussed in the following paragraph.

Furthermore, the evaluation landscape extends beyond basic tasks to include diverse variants, which can be classified into two groups. The first focuses on prompt manipulation, such as translation [28, 80, 101, 103], rephrasing, reverse query, and synonym substitution [30, 134]. The second designs structured scenarios, such as analogy completion or odd-one-out tasks [81]. Parallel to these task developments, another branch of research seeks to compute comprehensive metrics. To mitigate the reliance on point estimates, Scholten et al. [135] propose a probabilistic framework, which is calculated by extensively sampling the model generation. Other studies aggregate performance in numerous tasks through designed average scores [96, 143] or Cognitive Diagnosis Models (CDMs) [86].

4.1.2 Content. From the perspective of content, the unlearn set may originate from **real-world sources**, such as the Harry Potter series [41], or be **fictionally constructed**, as exemplified by the TOFU benchmark [105]. Real-world data exhibit richer content and more coherent logical relationships, thus being practically useful [80]. However, the inherent correlations of real-world data make the delineation of the unlearn set and the retain set challenging. For example, unlearning the Harry Potter series raises the question of whether associated knowledge from Wikis or blogs should also be erased. To address this issue, several studies employ fictional data generated via templates or LLMs [105, 172]. Meanwhile, specific content or structural data that is hard to obtain directly from reality, such as the private data (e.g., phone number, address) [126] or relationship graphs [123], can also be generated.

4.1.3 Experiment Paradigm. In the experiment, datasets can be broadly categorized into two classes based on whether fine-tuning is required. In the first category, the models perform unlearning **without fine-tuning** on the target dataset, which simplifies experimental setup. This includes the following scenarios. (1) The unlearn set is compiled directly from the model's pretraining data, such as subsets derived from the Pile [49]. (2) The data are manually verified to be present in the model, as in RWKU [80], ConceptVectors [62], and RETURN [96]. However, verifying the presence of facts in LLMs remains challenging and may affect reliability. (3) For security purposes, the model is required to erase certain knowledge regardless of its original presence, exemplified by WMDP [88] and UNCD [86]. In the second category, models are **first fine-tuned** on the full dataset before a subset is unlearned. This is essential when datasets are fictionally synthetic, such as TOFU [105], EDU-RELAT [172], and PISTOL [123], to ensure that the model acquires the target knowledge. Even for real-world corpora, fine-tuning helps to guarantee that the original model possesses knowledge of the unlearn set.

4.1.4 Existing Benchmarks and Datasets. A direct motivation for unlearning research comes from a number of works that aim to remove parts of the pretraining corpus [3, 17, 21, 72, 122, 149, 153, 162, 179]. Among these, the Pile dataset [49], which is commonly used in pretraining LLMs such as Pythia [11], is one of the most frequently adopted. Google Research [56] further introduced the Training Data Extraction Challenge (TDEC), a subset of 20,000 examples of The Pile that has been employed as an unlearn set in several studies. However, a major challenge is that the pretraining data for many state-of-the-art LLMs are not publicly available, and different models often use different corpora, significantly limiting the applicability of such datasets.

To address diverse research needs, numerous benchmarks and datasets have been developed, varying in motivation and application. Some focus on unlearning specific content, such as security information [86, 88], copyrighted material [41, 143], or private data [96, 105]. Others emphasize knowledge connectivity or semantic diversity to enhance unlearning robustness [62, 123, 172, 177]. Additionally, works such as [130, 156] explore continuous learning–unlearning settings. We summarize the characteristics of existing benchmarks in Table 4.

Unlearning evaluations also frequently adapt benchmarks from related fields such as model editing and LLM safety. These include CounterFactual [107] (used by [59, 118, 167]), PKU-SafeRLHF [74] (used by [75, 76, 188]), SQuAD [125] (used by [119, 130]), and ZsRE [181] (used by [160, 167]).

Class		Advantages	
Task	Cloze Continuation	(Free-form) simple data preparation.	Flexible position of questioning content. Long answer length.
	MCQA Short answer	(QA) clear questioning objectives.	Unique answer, easy for evaluation Various forms and scenarios.
Content	Real-world Fictional	Rich content, coherent logical relationships, practically useful. Easy to separate the unlearn/retain set, flexible to construct required content and format.	
Experiment	W/o fine-tuning W/ fine-tuning	Low computational cost and simple experiment. Ensure that the original model memorize the unlearn set, continuous learning-unlearning scenario.	

(a) Comparison of different tasks formats, data contents and experiment paradigms.

Benchmark	Real	Fic.	FT	Data (Free-form/QA)	Tasks	Used by
TOFU [105]		✓	✓	200×20 QA	SAQA	[21, 25, 47, 57, 67, 73, 75, 76, 81, 92, 95, 104, 106, 129, 134, 135, 139, 151, 154, 160, 163, 164, 175, 185, 188]
WMDP [88]	✓		✗	Papers & passages	3,668 MCQA	[10, 24, 35, 38, 43, 65, 75, 84, 92, 102, 129, 134, 154, 160]
WHP [41]	✓		✓	3.1M tokens	300 Conti + 30 Cloze	[25, 73, 75, 76, 103, 134, 135, 154, 160]
MUSE [143]	✓		✓	4.4M+6.5M tokens	Conti + SAQA	[78, 131, 160, 184, 197]
RWKU [80]	✓		✗	200 celebrities	3,268 cloze + 2,879 SAQA	[151, 184]
CoTaEval [168]	✓		✓	1K + 1K passages	1.5K Conti + 1K SAQA	[131]
KnowUnDo [155]	✓	✓	✓	2,649 QA	SAQA	[175]
PISTOL [123]		✓	✓	4 Graphs (50,95)	95 SAQA	[139]
LUME [126]	✓	✓	✓	1,387 documents	4,394 (Conti + SAQA)	SEMEval-2025 Task 4
ConceptVectors [62]	✓		✗	285×10 paragraph	285×10 Conti + 285×10 SAQA	-
EDU-RELAT [172]		✓	✓	700 QA	11×5 SAQA	-
ELUDe [30]	✓		✓	15,651+90,954 QA	MCQA + SAQA	-
FaithUn [177]	✓		✗	664 QA	8,377 MCQA	-
LLM Surgery [156]	✓	✓	✓	180K+1B tokens	24,800 MCQA	-
Restor [130]	✓		✓	3,000 passages	1,051 SAQA	-
RETURN [96]	✓		✗	2492×20 QA	SAQA	-
UNCD [86]	✓		✗	2.9M+3.3M tokens	36K MCQA	-
WPU [95]	✓		✗	100 people's Wiki	2,795 SAQA	-

(b) Outlines of benchmarks and datasets.

Table 4. Select a suitable benchmark. Part (a) organizes advantages of different tasks formats, data content and experiment paradigm. Part (b) outlines existing benchmarks and datasets with their data content (real or/and fictional, abbreviated as Fic.), experiment paradigm (with or without fine-tuning, abbreviated as FT), statistics of data (text/QA), evaluation tasks and applications in subsequent studies. The '+' in the data column distinguishes between the unlearn set and the retain set. The meaning of abbreviations: "SAQA": Short answer question & answer, "MCQA": Multiple choice QA.

4.2 Metrics

After applying a unlearning method to a model on a selected dataset, we need several suitable metrics to evaluate effectiveness. Recalling the goal of unlearning in Section 2.1.2, the first kind of metric examines the **knowledge memorization** (Section 4.2.1) of the unlearned model on the content of the unlearn set and the retain set. Due to the various capabilities of LLMs, such as language proficiency and reasoning ability, the unlearned model should still retain all the **model utility** (Section 4.2.2). Additionally, we expect the unlearning process is **robust** (Section 4.2.3) and **efficiency** (Section 4.2.4). Refer to Figure 6 for an illustration of different metrics.

4.2.1 Knowledge Memorization. In most cases, the ideal result of unlearning is expected to contain all of the retain set and none of the unlearn set. The first kind of metrics evaluates **knowledge memorization**, which examines whether certain data have been memorized in the model. Typically, the choice of knowledge memorization metric is tailored to the task format; for instance, accuracy is a direct measure for multiple choice questions. In general, metrics are categorized into two distinct classes according to their operational basis: those applied to the model's final outputs and those applied to the model's internal logits. We illustrate all the knowledge memorization metrics introduced in this section in Table 5.

Output-based. The most direct approach is have the model complete the selected task and compare the output with the ground truth. Given that a model's output may not perfectly align with ground-truth references, multiple metrics are employed to quantify the textual similarity between them. **Verbatim matching** represents the simplest and most computationally efficient approach, particularly suitable for short or categorical answers. For longer and more complex generations, studies such as [154, 167] relax the exact match criterion to require the strict inclusion of specific keywords in the outputs. This adaptation demonstrates strong performance on benchmarks like TOFU [105], where questions are often centered on a unique, identifiable entity (e.g., an author's name). Another way to relax is to check if there is a target token among the top-k tokens in the next token prediction process, and calculate metrics such as the top hit ratio (THR) [123]. **BLEU** [117] and **ROUGE** (primarily ROUGE-N and ROUGE-L) [89] are established NLP metrics that focus on precision and recall of n-gram or longest common subsequence (LCS), respectively. However, they treat all words with equal weight. To prioritize key information, Xu et al. [175] introduced the Entity Coverage Score (ECS), which extracts key entities using a model like deepseek-v3 and calculates similarity based solely on these entities.

In addition to the model-free metrics, some methods introduce external models for a better evaluation beyond lexical overlap. Among these model-based methods, **BertScore** [195] constitutes a major category, typically involving the conversion of text into embedding vectors and then the calculation of cosine similarity. Through this embedding transformation, the model can better handle semantic-level information, such as synonyms, negation words, and word order. For scenarios requiring more knowledge and understanding (such as recognizing that "born in London" and "born in the UK" are consistent), evaluation using **NLI models** can yield more accurate results. Furthermore, more universal evaluation methods include **human evaluation** or **external LLM assessments**. These methods are not only adaptable across diverse tasks, but can also comprehensively evaluate the wording and grammar of the outputs. However, they often function as black boxes and can be susceptible to biases inherent in human evaluators or proxy LLMs. Finally, several less prevalent metrics have also been applied in unlearning evaluations [57, 92], including METEOR [6], MAUVE [121], and Rep$_3$ [171].

Logit-based. The autoregressive nature of Large Language Models (LLMs) involves computing a probability distribution for the next token conditioned on the preceding sequence. This **next token probability** can thus serve as an indicator of the model's latent knowledge for a given prompt, where a higher probability signifies stronger retention of that specific token [41, 105]. A representative indicator calculated on this basis is perplexity, which is adopted as a metric in several studies [38, 73, 173, 179], under the premise that more firmly memorized knowledge typically yields

Obj.	Class	Name	Note (Advantages)	Used by
Output	Model-free	Verbatim	Simple, computationally efficient, strong performance on some benchmarks. → keyword [154, 167], THR [123], ES [20]	[21, 122, 126, 149, 151, 154, 155, 163, 164, 167]
		BLEU	Commonly used in translation (precision).	[5, 57, 66, 73, 92, 135, 174]
		ROUGE	Commonly used in text summary (recall). Include LCS, ROUGE-L and ROUGE-N. → ECS [175]	[21, 26, 30, 39, 57, 66, 67, 73, 75, 78, 81, 92, 95, 104–106, 118, 123, 126, 129, 131, 134, 135, 139, 160, 163, 165, 168, 169, 175, 184, 185, 188, 197]
	Model-based	BertScore [195]	Semantic information (synonym, etc.). → Calculated with other encoders	[92, 131, 134, 168, 169, 174, 175, 185]
		NLI	Knowledge and understanding.	[78, 96, 175, 185]
		Human eval	Adaptable across diverse tasks, comprehensive in multiple dimensions.	[91, 175]
		LLM eval	Similar to the notes of human eval.	[66, 86, 95, 106, 126, 134, 165, 175]
Logit	Prob.	Next token probability	Examination of the model's latent knowledge. e.g. $p(a \mid q)$ [105], perplexity → KL divergence [163], CI [106]	[38, 41, 73, 105, 106, 118, 163, 173, 179]
		Truth ratio [105]	Detect under- and over-unlearning at the distribution level.	[21, 30, 57, 67, 73, 75, 76, 92, 106, 118, 154, 160, 164, 188]
	Rank	Rank score	Uniform score distribution, easy for comparison. e.g. Exposure [19], MRR [85]	[3, 5, 123, 173]

Table 5. Statistics of the use of different knowledge memorization metrics in LLM unlearning evaluations. Red text marks the new method improved on the corresponding method. Blue text marks representative examples from the corresponding methods.

lower perplexity. A significant advancement is the **Truth Ratio** introduced by Maini et al. [105], which quantifies the likelihood of a correct answer relative to incorrect alternatives for a given question. Theoretically, a model lacking specific knowledge should exhibit a negligible difference in probability between correct and incorrect answers. The efficacy of unlearning can be further statistically validated by applying tests like the KS-Test to compare the distribution of Truth Ratios between the unlearned model and an expected baseline. The truth ratio can effectively detect under- and over-unlearning at the distribution level.

A known limitation of direct probability usage is the extreme variance in conditional probabilities across tokens, which can adversely affect metric stability. A straightforward mitigation is to utilize token **ranks** instead of raw probabilities. Sorting tokens by their probability in descending order and using the rank as the score results offer a more uniform score distribution [3]. This rank-based paradigm is also employed by several metrics adapted for unlearning evaluation [5, 123, 173]. For instance, *Exposure* [19], a key metric in memorization analysis, can be viewed as a rank-based variant of the Truth Ratio, substituting likelihood comparison with rank comparison. Similarly, the Mean Reciprocal Rank (MRR), prevalent in entity retrieval tasks [85], calculates the reciprocal average of the ranks of target tokens.

4.2.2 Model Utility. Beyond investigating model memorization, various methodologies are employed to assess the general utility of unlearned models. Some directly and efficiently computable indicators quantify specific aspects of model performance, which is referred to as **utility metrics**. Among these, Perplexity is one of the most frequently used measures; a lower perplexity signifies better fluency [91], higher model confidence [169], and improved meaningfulness of the generated content [92]. In addition to perplexity, numerous studies focus on lexical diversity, proposing

Name	Description	Used by
ARC [32]	Reasoning	[30, 31, 41, 57, 67, 73, 96, 154–156, 165, 167, 179, 185]
GSM8K [33]	Math	[3, 66, 167, 179, 185]
HellaSwag [186]	Commonsense	[30, 31, 41, 67, 73, 131, 154, 167]
MMLU [60]	Universal knowledge	[3, 24, 30, 31, 35, 38, 39, 43, 59, 65, 66, 80, 86, 88, 92, 102, 126, 129, 131, 134, 153–155, 165, 168, 179, 185, 197]
MT-Bench [200]	Multi-turn	[24, 38, 39, 86, 88, 131, 154, 168]
OpenBookQA [108]	Understanding & application	[30, 31, 41, 47, 67, 73, 154]
PIQA [12]	Commonsense	[30, 31, 41, 57, 73, 96]
TruthfulQA [90]	Mimic human Falsehoods	[31, 66, 80, 131, 155, 184, 185, 197]
WinoGrande [133]	Commonsense	[30, 31, 41, 67, 96, 131]

Table 6. Overview of general benchmarks widely used in unlearning evaluation.

metrics such as the mean number of distinct n-grams [91], the unique token ratio [92, 180], and token entropy [185]. As noted by Yuan et al. [185], reduced vocabulary diversity often correlates with token repetition in model outputs, indicating poorer readability and weaker overall utility. Also, established linguistic indices, including Brunet's Index [16] and Honore's Statistic [63], have been applied to assess lexical richness in unlearning contexts [175].

Meanwhile, comprehensive **general benchmarks** are commonly utilized to evaluate the overall performance of unlearning methods [80, 88]. Table 6 summarizes frequently adopted benchmarks and their usage statistics across different unlearning studies. Integrated evaluation frameworks and toolkits, such as Language Model Evaluation Harness [50], facilitate the systematic application of these benchmarks for LLM unlearning evaluation [3, 75, 155, 167].

4.2.3 Unlearning Robustness. Empirical studies indicate that many machine methods merely suppress the surface-level expression of specific knowledge, leaving the underlying representations vulnerable to various adversarial attacks [62, 102, 103]. To systematically assess robustness, a range of adversarial techniques from the security domain have been integrated into the evaluation of LLM unlearning [24, 62, 76, 80, 88, 103, 126, 134, 143], which we collectively refer to as attack techniques. Commonly adopted methods include: (1) **Attack on the input**, such as crafted jailbreak prompts [140], in-context learning adversarial attacks [170], GCG [203], AutoPrompt [144], BEAST [132], PAIR [23], persona modulation [138], JailbreakHub [141], many-shot jailbreaking [2]. (2) **Attack on the hidden layer**, such as probing techniques [1, 9], soft-prompt-based threats [24, 136], AnonAct [137], Logit Lens [114]. (3) **Membership inference attack (MIA)**, which is a privacy attack that determines whether specific data samples are part of a model's training set [145], including LOSS [182], Zlib Entropy [20], Min-K% Prob [142] and Min-K%++ Prob [193]. Further heuristic attacks have also been proposed for specific unlearning scenarios [3, 40, 139].

In response to the characteristics of LLM unlearning tasks, a relatively unique robustness evaluation method, relearning [81, 103], is also frequently used. **Relearning** evaluates an unlearned model by exposing it to a limited subset of the unlearned data. In in-context relearning, knowledge related to the unlearn set, such as book summaries or relevant background information, are included in the prompts when evaluating the unlearned LLM. When relearning by fine-tuning, the model is full-parameter or LoRA fine-tuned on a small portion of the unlearn set or a related set. Empirical studies consistently demonstrate that relearning can substantially degrade unlearning quality, causing the model to rapidly recapitulate a significant portion of unlearned knowledge from sparse cues [59, 81, 87, 103, 187], or begin to systematically avoid generating content related to the unlearning target, even when contextually prompted [81]. More critically, Doshi and Stickland [38] reveal that fine-tuning on an entirely benign dataset can also reverse the unlearning effects, restoring model performance to a level comparable to its state before unlearning.

4.2.4 Unlearning efficiency. While the majority of existing studies concentrate on the efficacy of unlearning, the resource overhead of deploying such algorithms under real-world constraints remains a crucial consideration, including both memory occupation [7, 76] and computational time consumption. For time cost, a straightforward approach is to directly measure the algorithm's runtime during experiments [7, 26, 31, 76]. Since computational speed is closely tied to GPU performance, some studies convert raw runtime into GPU hours (i.e., number of GPUs × training hours) to facilitate comparison [73, 129]. Nonetheless, fair cross-study comparisons remain challenging due to variations in experimental environments. The most reliable method is to execute all algorithms under controlled conditions, though this is often resource-intensive. Alternatively, several works estimate time consumption theoretically, using metrics such as floating-point operations [21] or gradient computation budgets [169]. However, discrepancies between theoretical estimates and actual runtime may arise due to differences in implementation and hardware optimization.

5 Challenges and Future Directions

5.1 Challenges

5.1.1 Definition and Evaluation of Unlearning. In Section 2.1, we characterize the goal of LLM unlearning as ensuring that "the unlearned model should no longer memorize information from the unlearn set while preserving all other knowledge." However, two key issues remain ambiguous, leading to divergent definitions of unlearning across the literature and, consequently, to inconsistent and imprecise evaluation practices.

(1) How should memorization be defined and detected? Most studies assess memorization based on model output in specific tasks, yet disagree on the criteria for judging these outputs. For the content related to the unlearn set, some works argue that the model should simply avoid generating such content [41], while others require it to explicitly respond with "I don't know" [139]. Another line of research proposes that the unlearned model should produce outputs similar to those of a hypothetical retrained model, such as giving a specific incorrect answer [130]. When direct output is insufficient, adversarial methods are sometimes employed to expose memorization. However, such approaches face inherent limitations: Overly weak attacks may fail to detect memorization, whereas overly strong ones can force the model to generate arbitrary content, casting doubt on their reliability as auditing tools [25]. Moreover, certain attack methods are considered ill-suited to the LLM context, such as MIA, which typically requires training numerous shadow models, thus being both data-prohibitive and computationally intractable for LLMs [92].

(2) What should constitute the unlearn set? For synthetic datasets such as TOFU [105], this question is relatively straightforward. However, in real-world scenarios, data interconnectivity complicate the identification of appropriate unlearning targets. Tian et al. [155] adopt a legal perspective to determine which copyrighted or private data should be unlearned, while Wu et al. [172] construct relationship graphs to identify necessary data for removal. Despite their merits, these methods remain reliant on manual, domain-specific analysis, and lack generalizability.

5.1.2 Effect of Unlearning. Unlearning has different effects on different languages and data, further increasing the difficulty of designing and evaluating unlearning algorithms.

Effects across languages. Some studies conduct evaluations with prompts translated into languages other than English, finding that monolingual unlearning is fundamentally insufficient for multilingual LLMs [80, 103]. Furthermore, more languages that systematically divide into high- and low-resource are used in evaluations [28, 101], revealing the fact that unlearning in one language does not necessarily transfer to others and could even inadvertently reinforce harmful content across languages. Together, these findings underscore a critical consensus: effective and secure unlearning necessitates multilingual joint unlearning strategies that are designed to address model behavior holistically in all languages.

Effects across data. From the perspective of data distribution, Baluta et al. [5] demonstrate that out-of-distribution (OOD) data require more gradient ascent but offer a better unlearning quality, whereas in-distribution data allow faster unlearning but severely compromise model utility, illustrating a fundamental trade-off between unlearning efficiency and model preservation. Considering the logical connectivity of the data, Choi et al. [29] identify that current unlearning methods struggle with multi-hop knowledge, where unlearning one intermediate fact in a chain often fails to remove the entire logical sequence. Furthermore, some studies investigate the impact on adjacent data after performing unlearning on selected data, identifying phenomena called "transfer unlearning" [100], "ripple effect" [196] and "onion effect" [14]. These effects highlight the intricate and unpredictable consequences of unlearning, emphasizing the need for careful monitoring to ensure that unlearning achieves its intended goals without introducing new risks.

5.1.3 Unlearning in Reality. A significant challenge lies in the **scaling gap** between experimental settings and real-world conditions. Current unlearning experiments are largely limited to models with fewer than 10 billion parameters and unlearning sets under 1 billion instances, raising concerns about the applicability of these methods to larger models and datasets. Shi et al. [143] analyze how evaluation metrics evolve as the size of the unlearn set increases, providing insight into scalability. On the other hand, in practical deployments, large models are often compressed, such as using quantization for efficiency. Notably, Zhang et al. [197] demonstrate that quantizing unlearned models can inadvertently reactivate unlearned knowledge, highlighting a key scalability challenge.

In commercial applications, unlearning requests typically arrive sequentially, requiring models to **continuously unlearning** while maintaining performance [130, 156]. To assess long-term viability, Shi et al. [143] collect model checkpoints after processing each sequential request and track evaluation metrics over time. This approach helps quantify the cumulative impact of repeated unlearning and the model's ability to sustain utility. Unfortunately, current unlearning methods are not yet ready to handle sequential unlearning.

5.2 Future Directions

5.2.1 Unlearning in Specialized Architectures and Scenarios. The field is moving towards addressing unlearning in sophisticated model architectures. Cheng and Amiri [27] pioneer this effort for tool-augmented large language models (LLMs) by proposing ToolDelete, the first unlearning framework designed to remove a specific "skill" or the ability to use a particular tool, and they introduce a new membership inference attack (MIA) for evaluation. Similarly, the unique structure of Mixture-of-Experts (MoE) models presents a distinct challenge. Zhuang et al. [202] find that unlearning a single expert is insufficient and propose the Selected Expert Unlearning Framework (SEUF) to effectively perform unlearning on MoE models. These works demonstrate that effective unlearning requires bespoke algorithms tailored to a model's specific architecture and knowledge organization.

5.2.2 Unlearning as Tools. Unlearning is not only a goal in itself, but also a powerful tool when we further expand the scope of unlearning targets. Firstly, when choosing injected trojans or backdoor triggers as the target, unlearning can be an effective tool in defense [61, 77, 198]. Similarly, an opposite target, such as removing the safety alignment or disrupting the subsequent fine-tuning process on a base model, can convert unlearning to a means of attack [127]. Furthermore, if unlearning the selected training data and examining the changes in the model before and after unlearning, we can have novel insights into how different data components contribute to and influence the final model capabilities [70, 199]. A powerful and accurate unlearning method will play an important role as a tool.

5.2.3 Unlearning beyond Data. Most existing studies focus on unlearning specific data instances. However, in practical scenarios, unlearning requests often target not only concrete data but also abstract concepts or capabilities, such as erroneous reasoning patterns, harmful ethical values, or unsafe skills [86, 88]. Extending unlearning beyond the data to encompass abstract constructs is essential to prevent the propagation of incorrect or harmful knowledge. Achieving this goal may present two main pathways: one is to precisely identify and modify parameters or representations associated with particular concepts or abilities; the other leverages established alignment techniques, such as reinforcement learning, by designing appropriate reward mechanisms that penalize the generation of undesirable content.

5.2.4 Robust Unlearning. In light of the observed fragility of LLM unlearning, a significant research direction aims to develop techniques that enhance its robustness and long-term stability. These defensive efforts pursue two primary objectives: (1) to ensure that knowledge removal is thorough and persistent, thereby resisting attempts at recovery; and (2) to prevent the unlearning procedure from introducing new vulnerabilities or unintended side effects. Several existing studies address the first objective through robust unlearning frameworks [42, 68, 152, 191] or methods that strengthen the robustness of unlearned models [68, 191]. Nevertheless, given the proliferation of advanced attacks, achieving truly robust unlearning remains a critical and ongoing topic.

5.2.5 Verifiable and Certifiable Unlearning. In most current practices, unlearning is applied to models that have already internalized the content targeted for removal through opaque mechanisms, complicating the certification of unlearning effectiveness. However, from legal, safety, and social trust perspectives, achieving verifiable and trustworthy unlearning remains critically important. To validate existing unlearning methods, it is essential to establish a fair and comprehensive evaluation benchmark. Looking ahead, future work may also draw inspiration from frameworks such as SISA by designing structured data storage and training protocols to enable intrinsically verifiable unlearning.

6 Conclusions

Machine unlearning has emerged as a pivotal technique to address critical challenges in large language models, including privacy protection, copyright compliance, and safety enhancement. In this survey, we provide a comprehensive review of work dedicated to LLM unlearning, including the definition and goal of LLM unlearning, the most recent LLM unlearning methods, and commonly used datasets and evaluation metrics of unlearning. Despite significant progress, the field of LLM unlearning remains in its early stages, with fundamental challenges in the definition, evaluation, effects and practical deployment of unlearning. Furthermore, we suggest several promising directions for future research. We hope that this survey can provide readers with a general understanding of recent progress in this field and shed some light on future developments.

Acknowledgments

This work is partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0680201), the Beijing Natural Science Foundation (QY23156, 4252023).

References

[1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644.*

[2] Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Fracesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanhan, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. 2025. Many-shot jailbreaking. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

[3] Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. 2025. REVS: Unlearning sensitive information in language models via rank editing in the vocabulary space. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14774–14797, Vienna, Austria. Association for Computational Linguistics.

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862.*

[5] Teodora Baluta, Pascal Lamblin, Daniel Tarlow, Fabian Pedregosa, and Gintare Karolina Dziugaite. 2024. Unlearning in- vs. out-of-distribution data in LLMs under gradient-based methods. In *Neurips Safe Generative AI Workshop 2024.*

[6] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

[7] Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2023. Privacy adhering machine un-learning in NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 268–277, Nusa Dua, Bali. Association for Computational Linguistics.

[8] 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. 2025. Open Problems in Machine Unlearning for AI Safety. ArXiv:2501.04952 [cs].

[9] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

[10] Karuna Bhaila, Minh-Hao Van, and Xintao Wu. 2025. Soft prompting for unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4046–4056, Albuquerque, New Mexico. Association for Computational Linguistics.

[11] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

[12] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):7432–7439.

[13] Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review*, 58(3):90.

[14] Jaydeep Borkar. 2023. What can we learn from data leakage and unlearning for law? In *Proceedings of the 1st Workshop on Generative AI and Law (co-located with ICML 2023).* Accepted workshop paper.

[15] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

[16] Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution.* Slatkine.

[17] George-Octavian Bărbulescu and Peter Triantafillou. 2024. To each (textual sequence) its own: improving memorized-data unlearning in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

[18] Yinzhi Cao and Junfeng Yang. 2015. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480.

[19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

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

[21] Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2024. Towards robust and cost-efficient knowledge unlearning for large language models. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

[22] Hwan Chang and Hwanhee Lee. 2025. Which Retain Set Matters for LLM Unlearning? A Case Study on Entity Unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5966–5982. Association for Computational Linguistics.

[23] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

[24] Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, et al. 2025. Model tampering attacks enable more rigorous evaluations of llm capabilities. *arXiv preprint arXiv:2502.05209*.

[25] Haokun Chen, Sebastian Szyller, Weilin Xu, and Nageen Himayat. 2025. Soft token attacks cannot reliably audit unlearning in large language models. *arXiv preprint arXiv:2502.15836*.

[26] Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics.

[27] Jiali Cheng and Hadi Amiri. 2025. Tool unlearning for tool-augmented LLMs. In *Forty-second International Conference on Machine Learning*.

[28] Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. Cross-lingual unlearning of selective knowledge in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10732–10747, Miami, Florida, USA. Association for Computational Linguistics.

[29] Minseok Choi, ChaeHun Park, Dohyun Lee, and Jaegul Choo. 2024. Breaking chains: Unraveling the links in multi-hop knowledge unlearning.

[30] Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2025. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28280–28297, Vienna, Austria. Association for Computational Linguistics.

[31] Somnath Basu Roy Chowdhury, Krzysztof Marcin Choromanski, Arijit Sehanobish, Kumar Avinava Dubey, and Snigdha Chaturvedi. 2025. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

[32] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

[33] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[34] A Feder Cooper, Christopher A Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, et al. 2024. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice. *arXiv preprint arXiv:2412.06966*.

[35] Huu-Tien Dang, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. 2025. On effects of steering latent representation for large language model unlearning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23733–23742.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[37] Omkar Dige, Diljot Singh, Tsz Fung Yau, Qixuan Zhang, Borna Bolandraftar, Xiaodan Zhu, and Faiza Khan Khattak. 2024. Mitigating Social Biases in Language Models through Unlearning. *arXiv preprint 2406.13551*.

[38] Jai Doshi and Asa Cooper Stickland. 2024. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*.

[39] Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2025. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5176–5200, Albuquerque, New Mexico. Association for Computational Linguistics.

[40] Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. 2025. Textual unlearning gives a false sense of unlearning. In *Forty-second International Conference on Machine Learning*.

[41] Ronen Eldan and Mark Russinovich. 2024. Who's harry potter? approximate unlearning for LLMs.

[42] Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. 2025. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *Forty-second International Conference on Machine Learning*.

[43] Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. 2024. Applying sparse autoencoders to unlearn knowledge in language models. In *Neurips Safe Generative AI Workshop 2024*.

[44] XiaoHua Feng, Chaochao Chen, Yuyuan Li, and Zibin Lin. 2024. Fine-grained Pluggable Gradient Ascent for Knowledge Unlearning in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10141–10155. Association for Computational Linguistics.

[45] Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. Unlearning climate misinformation in large language models. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 178–192, Bangkok, Thailand. Association for Computational Linguistics.

[46] Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

[47] Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. 2025. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*.

[48] Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. 2024. Ethos: Rectifying language models in orthogonal parameter space. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2054–2068, Mexico City, Mexico. Association for Computational Linguistics.

[49] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

[50] Leo Gao, Jonathan Tow, Stella Biderman, Shawn Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jasmine Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2021. A framework for few-shot language model evaluation.

[51] Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*.

[52] Ruotong Geng, Mingyang Geng, Shangwen Wang, Haotian Wang, Zhipeng Lin, and Dezun Dong. 2025. Mitigating sensitive information leakage in llms4code through machine unlearning. *arXiv preprint arXiv:2502.05739*.

[53] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

[54] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

[55] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

[56] Google Research. 2023. Language model extraction benchmark. https://github.com/google-research/lm-extraction-benchmark. Accessed: 2025-04-10.

[57] Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. 2024. MEOW: MEMOry Supervised LLM Unlearning Via Inverted Facts. *arXiv preprint 2409.11844*.

[58] Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2024. Robust Unlearning via Mechanistic Localizations. In *ICML 2024 Workshop on Mechanistic Interpretability*.

[59] Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2025. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*.

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

[61] Adriano Hernandez. 2024. If you don't understand it, don't use it: Eliminating trojans with filters between layers. *arXiv preprint arXiv:2407.06411*.

[62] Yihuai Hong, Lei Yu, Shauli Ravfogel, Haiqin Yang, and Mor Geva. 2024. Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces. *arXiv preprint 2406.11614*.

[63] Antony Honoré et al. 1979. Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, 7(2):172–177.

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

[65] Jinwei Hu, Zhenglin Huang, Xiangyu Yin, Wenjie Ruan, Guangliang Cheng, Yi Dong, and Xiaowei Huang. 2025. Falcon: Fine-grained activation manipulation by contrastive orthogonal unalignment for large language model. *arXiv*

preprint arXiv:2502.01472.

[66] Xinshuo Hu, Dongfang Li, Baotian Hu, Zihao Zheng, Zhenyu Liu, and Min Zhang. 2024. Separate the wheat from the chaff: Model deficiency unlearning via parameter-efficient module operation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18252–18260.

[67] James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2025. Offset unlearning for large language models. *Transactions on Machine Learning Research*.

[68] Dang Huu-Tien, Hoang Thanh-Tung, Anh Bui, Le-Minh Nguyen, and Naoya Inoue. 2025. Improving llm unlearning robustness via random perturbations. *arXiv preprint arXiv:2501.19202*.

[69] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

[70] Masaru Isonuma and Ivan Titov. 2024. Unlearning traces the influential training data of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6312–6325, Bangkok, Thailand. Association for Computational Linguistics.

[71] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge Unlearning for Mitigating Privacy Risks in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

[72] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

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

[74] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: towards improved safety alignment of llm via a human-preference dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[75] Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. 2025. Wagle: strategic weight attribution for effective and modular unlearning in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

[76] Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA. Association for Computational Linguistics.

[77] Peihai Jiang, Xixiang Lyu, Yige Li, and Jing Ma. 2025. Backdoor token unlearning: Exposing and defending backdoors in pretrained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24285–24293.

[78] Weipeng Jiang, Juan Zhai, Shiqing Ma, Ziyan Lei, Xiaofei Xie, Yige Wang, and Chao Shen. 2025. Holistic audit dataset generation for llm unlearning via knowledge graph traversal and redundancy removal. *arXiv preprint arXiv:2502.18810*.

[79] Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. 2025. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11278–11294, Albuquerque, New Mexico. Association for Computational Linguistics.

[80] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263.

[81] Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. Towards robust evaluation of unlearning in LLMs via data transformations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12100–12119, Miami, Florida, USA. Association for Computational Linguistics.

[82] Dahyun Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park, and Heuiseok Lim. 2025. CoME: An unlearning-based approach to conflict-free model editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6410–6422, Albuquerque, New Mexico. Association for Computational Linguistics.

[83] Swanand Kadhe, Anisa Halimi, Ambrish Rawat, and Nathalie Baracaldo. 2023. FairSISA: Ensemble post-processing to improve fairness of unlearning in LLMs. In *Socially Responsible Language Modelling Research*.

[84] Arinbjörn Kolbeinsson, Kyle O'Brien, Tianjin Huang, Shanghua Gao, Shiwei Liu, Jonathan Richard Schwarz, Anurag Jayant Vaidya, Faisal Mahmood, Marinka Zitnik, Tianlong Chen, and Thomas Hartvigsen. 2025. Composable interventions for language models. In *The Thirteenth International Conference on Learning Representations*.

[85] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*, pages 2863–2872. PMLR.

[86] Yicheng Lang, Kehan Guo, Yue Huang, Yujun Zhou, Haomin Zhuang, Tianyu Yang, Yao Su, and Xiangliang Zhang. 2025. Beyond single-value metrics: Evaluating and enhancing llm unlearning with cognitive diagnosis. *arXiv preprint arXiv:2502.13996*.

[87] Simon Lermen and Charlie Rogers-Smith. 2024. LoRA fine-tuning efficiently undoes safety training in llama 2-chat 70b. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

[88] 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. 2024. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550.

[89] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[90] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

[91] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

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

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

[94] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[95] Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. 2024. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8708–8731, Miami, Florida, USA. Association for Computational Linguistics.

[96] Zhenhua Liu, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. 2025. Learning to refuse: Towards mitigating privacy risks in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1683–1698, Abu Dhabi, UAE. Association for Computational Linguistics.

[97] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Machine Unlearning in Generative AI: A Survey. *arXiv preprint 2407.20516*.

[98] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, Bangkok, Thailand. Association for Computational Linguistics.

[99] Tyler Lizzo and Larry Heck. 2025. UNLEARN efficient removal of knowledge in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7257–7268, Albuquerque, New Mexico. Association for Computational Linguistics.

[100] Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2024. Towards transfer unlearning: Empirical evidence of cross-domain bias mitigation.

[101] Taiming Lu and Philipp Koehn. 2025. Learn and unlearn: Addressing misinformation in multilingual llms.

[102] Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2025. An adversarial perspective on machine unlearning for AI safety. *Transactions on Machine Learning Research*.

[103] Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight Methods to Evaluate Robust Unlearning in LLMs. *arXiv preprint 2402.16835*.

[104] Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. 2025. Unveiling entity-level unlearning for large language models: A comprehensive analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5345–5363, Abu Dhabi, UAE. Association for Computational Linguistics.

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

[106] Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. 2025. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752, Abu Dhabi, UAE. Association for Computational Linguistics.

[107] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

[108] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

[109] Andrei Ioan Muresanu, Anvith Thudi, Michael R. Zhang, and Nicolas Papernot. 2025. Fast exact unlearning for in-context learning data for LLMs. In *Forty-second International Conference on Machine Learning*.

[110] Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. Blog post on mechanistic interpretability. Accessed: 2025-10-28.

[111] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

[112] Andrew Ng. 2011. Sparse autoencoder. CS294A Lecture Notes, Stanford University. Accessed: 2025-10-28.

[113] Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2024. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5716–5731, Bangkok, Thailand. Association for Computational Linguistics.

[114] nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong blog post. Accessed: 2025-10-28.

[115] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

[116] Zibin Pan, Shuwen Zhang, Yuesheng Zheng, Chi Li, Yuheng Cheng, and Junhua Zhao. 2025. Multi-objective large language model unlearning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[117] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[118] Vaidehi Patil, Elias Stengel-Eskin, and Mohit Bansal. 2025. UPCORE: Utility-Preserving Coreset Selection for Balanced Unlearning. *arXiv preprint 2502.15082*.

[119] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40034–40050. PMLR.

[120] Joshua Peterson, Stephan Meylan, and David Bourgin. 2019. Open clone of openai's unreleased webtext dataset scraper.

[121] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

[122] Nicholas Pochinkov and Nandi Schoots. 2024. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*.

[123] Xinchi Qiu, William F. Shen, Yihong Chen, Nicola Cancedda, Pontus Stenetorp, and Nicholas D. Lane. 2024. PISTOL: Dataset Compilation Pipeline for Structural Unlearning of LLMs. In *Proceedings of the GENAI Evaluation Workshop at KDD 2024*, Barcelona, Spain.

[124] Youyang Qu, Ming Ding, Nan Sun, Kanchana Thilakarathna, Tianqing Zhu, and Dusit Niyato. 2025. The Frontier of Data Erasure: A Survey on Machine Unlearning for Large Language Models . *Computer*, 58(01):45–57.

[125] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

[126] Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025. Lume: Llm unlearning with multitask evaluations. *arXiv preprint arXiv:2502.15097*.

[127] Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Ye Wang, and Shagufta Mehnaz. 2025. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):20139–20147.

[128] Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–13.

[129] Jie Ren, Zhenwei Dai, Xianfeng Tang, Hui Liu, Jingying Zeng, Zhen Li, Rahul Goutam, Suhang Wang, Yue Xing, and Qi He. 2025. A general framework to enhance fine-tuning-based llm unlearning. *arXiv preprint arXiv:2502.17823*.

[130] Keivan Rezaei, Khyathi Chandu, Soheil Feizi, Yejin Choi, Faeze Brahman, and Abhilasha Ravichander. 2024. Restor: Knowledge recovery through machine unlearning. *arXiv preprint arXiv:2411.00204*.

[131] Mark Russinovich and Ahmed Salem. 2025. Obliviate: Efficient unmemorization for protecting intellectual property in large language models. *arXiv preprint arXiv:2502.15010*.

[132] Vinu Sankar Sadasivan, Shoumik Saha, Gaurang Sriramanan, Priyatham Kattakinda, Atoosa Chegini, and Soheil Feizi. 2024. Fast adversarial attacks on language models in one GPU minute. In *Forty-first International Conference on Machine Learning*.

[133] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

[134] Debdeep Sanyal and Murari Mandal. 2025. Agents are all you need for LLM unlearning. In *Second Conference on Language Modeling*.

[135] Yan Scholten, Stephan Günnemann, and Leo Schwinn. 2025. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*.

[136] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Günnemann. 2024. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *Advances in Neural Information Processing Systems*, 37:9086–9116.

[137] Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. 2024. Extracting unlearned information from LLMs with activation steering. In *Neurips Safe Generative AI Workshop 2024*.

[138] Rusheb Shah, Quentin Feuillade Montixi, Soroush Pour, Arush Tagade, and Javier Rando. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. In *Socially Responsible Language Modelling Research*.

[139] William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. 2025. Lunar: Llm unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*.

[140] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.

[141] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 1671–1685, New York, NY, USA. Association for Computing Machinery.

[142] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

[143] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

[144] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

[145] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

[146] Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge Unlearning for LLMs: Tasks, Methods, and Challenges. *arXiv preprint 2311.15766*.

[147] Naman Deep Singh, Maximilian Müller, Francesco Croce, and Matthias Hein. 2025. Unlearning That Lasts: Utility-Preserving, Robust, and Almost Irreversible Forgetting in LLMs. *arXiv preprint 2509.02820*.

[148] Yash Sinha, Murari Mandal, and Mohan Kankanhalli. 2025. UnSTAR: Unlearning with self-taught anti-sample reasoning for LLMs. *Transactions on Machine Learning Research*.

[149] Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

[150] Chen Sun, Nolan Andrew Miller, Andrey Zhmoginov, Max Vladymyrov, and Mark Sandler. 2024. Learning and unlearning of fabricated knowledge in language models. In *ICML 2024 Workshop on Mechanistic Interpretability*.

[151] Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. *arXiv preprint arXiv:2410.00382*.

[152] Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. 2025. Tamper-resistant safeguards for open-weight LLMs. In *The Thirteenth International Conference on Learning Representations*.

[153] Rishub Tamirisa, Bhrugu Bharathi, Andy Zhou, Bo Li, and Mantas Mazeika. 2024. Toward Robust Unlearning for LLMs. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

[154] Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail Baselines for Unlearning in LLMs. *arXiv preprint 2403.03329*.

[155] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1524–1537, Miami, Florida, USA. Association for Computational Linguistics.

[156] Akshaj Kumar Veldanda, Shi-Xiong Zhang, Anirban Das, Supriyo Chakraborty, Stephen Rawls, Sambit Sahu, and Milind Naphade. 2024. Llm surgery: Efficient knowledge unlearning and editing in large language models. *arXiv preprint arXiv:2409.13054*.

[157] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

[158] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[159] Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2025. Balancing Forget Quality and Model Utility: A Reverse KL-Divergence Knowledge Distillation Approach for Better Unlearning in LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1306–1321. Association for Computational Linguistics.

[160] Huazheng Wang, Yongcheng Jing, Haifeng Sun, Yingjie Wang, Jingyu Wang, Jianxin Liao, and Dacheng Tao. 2025. Erasing without remembering: Implicit knowledge forgetting in large language models. *arXiv preprint arXiv:2502.19982*.

[161] 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13264–13276.

[162] Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):843–851.

[163] Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. 2025. Towards effective evaluations and comparisons for LLM unlearning methods. In *The Thirteenth International Conference on Learning Representations*.

[164] Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q. Weinberger. 2024. Rethinking LLM Unlearning Objectives: A Gradient Perspective and Go Beyond. In *The Thirteenth International Conference on Learning Representations*.

[165] Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. 2025. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *IEEE Transactions on Dependable and Secure Computing*.

[166] Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Shah, Yujia Bao, Yang Liu, and Wei Wei. 2025. LLM unlearning via loss adjustment with only forget data. In *The Thirteenth International Conference on Learning Representations*.

[167] Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. 2025. Large scale knowledge washing. In *The Thirteenth International Conference on Learning Representations*.

[168] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2025. Evaluating copyright takedown methods for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

[169] Rongzhe Wei, Mufei Li, Mohsen Ghassemi, Eleonora Kreacic, Yifan Li, Xiang Yue, Bo Li, Vamsi K. Potluru, Pan Li, and Eli Chien. 2025. Underestimated privacy risks for minority populations in large language model unlearning. In *Forty-second International Conference on Machine Learning*.

[170] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2023. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*.

[171] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

[172] Ruihan Wu, Chhavi Yadav, Ruslan Salakhutdinov, and Kamalika Chaudhuri. 2025. Evaluating deep unlearning in large language models. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*.

[173] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and Editing Privacy Neurons in Pretrained Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886.

[174] YuXuan Wu, Bonaventure F. P. Dossou, and Dianbo Liu. 2024. Codeunlearn: Amortized zero-shot machine unlearning in language models using discrete concept. In *Neurips Safe Generative AI Workshop 2024*.

[175] Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. 2025. ReLearn: Unlearning via learning for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5967–5987, Vienna, Austria. Association for Computational Linguistics.

[176] Yi Xu. 2024. Machine Unlearning for Traditional Models and Large Language Models: A Short Survey. *arXiv preprint 2404.01206*.

[177] Nakyeong Yang, Minsung Kim, Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2025. Faithun: Toward faithful forgetting in language models by investigating the interconnectedness of knowledge. In *Submitted to ACL Rolling Review - February 2025*. Under review.

[178] Zhou Yang and David Lo. 2024. Hotfixing large language models for code. *arXiv preprint arXiv:2408.05727*.

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

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

[181] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240.

[182] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

[183] 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*, pages 6032–6048.

[184] Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25769–25777.

[185] Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2025. A closer look at machine unlearning for large language models. In *The Thirteenth International Conference on Learning Representations*.

[186] 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*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

[187] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2024. Removing RLHF protections in GPT-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 681–687, Mexico City, Mexico. Association for Computational Linguistics.

[188] Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, and Haifeng Chen. 2025. Resolving editing-unlearning conflicts: A knowledge codebook framework for large language model updating. *arXiv preprint arXiv:2502.00158*.

[189] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramer, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 39321–39362. Curran Associates, Inc.

[190] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2025. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, 5(3):2445–2454.

[191] Eric Zhang, Leshem Choshen, and Jacob Andreas. 2024. Unforgettable generalization in language models. In *First Conference on Language Modeling*.

[192] Jinghan Zhang, shiqi chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operation. In *Advances in Neural Information Processing Systems*, volume 36, pages 12589–12610. Curran Associates,

Inc.

[193] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

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

[195] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

[196] Zhexin Zhang, Junxiao Yang, Yida Lu, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2025. From theft to bomb-making: The ripple effect of unlearning in defending against jailbreak attacks.

[197] Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations*.

[198] Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Anh Tuan Luu. 2025. Unlearning backdoor attacks for LLMs with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4937–4952, Vienna, Austria. Association for Computational Linguistics.

[199] Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Zhouhao Sun, Shi Jun, Ting Liu, and Bing Qin. 2024. Deciphering the impact of pretraining data on large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9386–9406, Bangkok, Thailand. Association for Computational Linguistics.

[200] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[201] Shiji Zhou, Lianzhe Wang, Jiangnan Ye, Yongliang Wu, and Heng Chang. 2024. On the limitations and prospects of machine unlearning for generative ai. *arXiv preprint arXiv:2408.00376*.

[202] Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, and Xiangliang Zhang. 2025. SEUF: Is unlearning one expert enough for mixture-of-experts LLMs? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8664–8678, Vienna, Austria. Association for Computational Linguistics.

[203] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.