Rethinking Machine Unlearning for Large Language Models

Sijia Liu[1,2] **Yuanshun Yao**[3†] **Jinghan Jia**[1†] **Stephen Casper**[4]
Nathalie Baracaldo[2] **Peter Hase**[5] **Yuguang Yao**[1] **Chris Yuhao Liu**[6]
Xiaojun Xu[3] **Hang Li**[3] **Kush R. Varshney**[2] **Mohit Bansal**[5] **Sanmi Koyejo**[7] **Yang Liu**[3,6]

[1]Computer Science & Engineering Dept., Michigan State University, USA
[2]IBM Research, USA
[3]ByteDance Research, USA
[4]Computer Science and Artificial Intelligence Laboratory, MIT, USA
[5]Computer Science Dept., University of North Carolina at Chapel Hill, USA
[6]Computer Science & Engineering Dept., University of California, Santa Cruz, USA
[7]Computer Science Dept., Stanford University, USA
[†]These authors contributed equally to this work.

Abstract

We explore machine unlearning (MU) in the domain of large language models (LLMs), referred to as LLM unlearning. This initiative aims to eliminate undesirable data influence (*e.g.*, sensitive or illegal information) and the associated model capabilities, while maintaining the integrity of essential knowledge generation and not affecting causally unrelated information. We envision LLM unlearning becoming a pivotal element in the life-cycle management of LLMs, potentially standing as an essential foundation for developing generative AI that is not only safe, secure, and trustworthy, but also resource-efficient without the need of full retraining. We navigate the unlearning landscape in LLMs from conceptual formulation, methodologies, metrics, and applications. In particular, we highlight the often-overlooked aspects of existing LLM unlearning research, *e.g.*, unlearning scope, data-model interaction, and multifaceted efficacy assessment. We also draw connections between LLM unlearning and related areas such as model editing, influence functions, model explanation, adversarial training, and reinforcement learning. Furthermore, we outline an effective assessment framework for LLM unlearning and explore its applications in copyright and privacy safeguards and sociotechnical harm reduction.

1 Introduction

Large language models (LLMs) have shown exceptional proficiency in generating text that closely resembles human-authored content. However, their ability to memorize extensive corpora may also lead to ethical and security concerns. These include societal biases and stereotyping (Bender et al., 2021; Motoki et al., 2023; Kotek et al., 2023), the generation of sensitive, private, harmful, or illegal content (Nasr et al., 2023; Wen et al., 2023; Karamolegkou et al., 2023; Patil et al., 2024), ease of jailbreaking (Wei et al., 2023; Zou et al., 2023; Liu et al., 2023b), and possible malicious use in developing cyberattacks or bioweapons (Barrett et al., 2023; Hendrycks et al., 2023; Li et al., 2024a). These concerns emphasize the need to adeptly and efficiently tailor pre-trained LLMs to suit diverse safety contexts while meeting specific requirements of users and sectors.

With the costly and prolonged training periods of LLMs, retraining these models to eliminate undesirable data effects is often impractical (Brown et al., 2020; Yao et al., 2024). Machine unlearning (MU) has emerged as an alternative to remove the influence of undesirable data and associated model capabilities from the pre-trained models (Cao & Yang, 2015; Bourtoule et al., 2021; Nguyen et al., 2022; Si et al., 2023; Zhang et al., 2023a; Eldan & Russinovich, 2023;

Correspondence to: Sijia Liu (`liusiji5@msu.edu`), Yang Liu (`yangliu@ucsc.edu`).

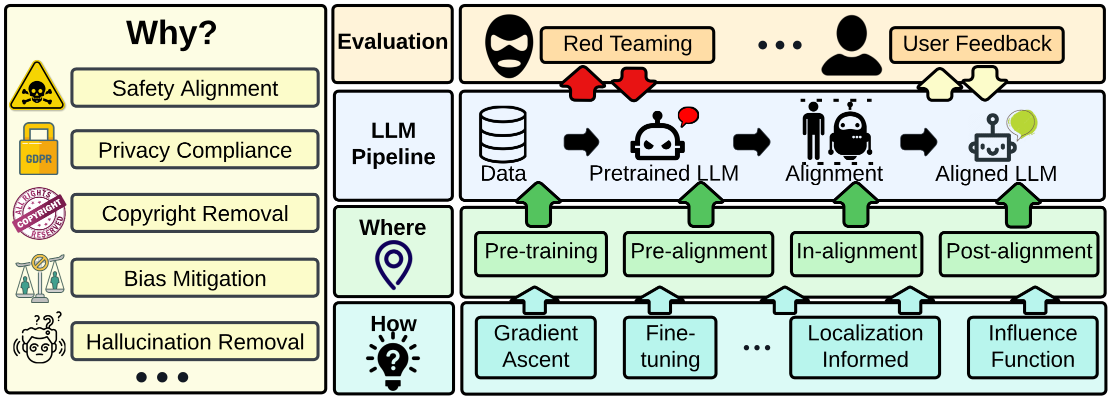

Figure 1: Demonstration of how MU can be incorporated into LLM development cycle. The landscape of LLM unlearning will be mainly navigated from applications ('why'), methods ('where' and 'how'), and evaluations.

Yao et al., 2023). For example, MU is used as a strategy to prevent the generation of copyrighted material from the Harry Potter series (Eldan & Russinovich, 2023). In the context of classification tasks, MU has been extensively studied (Ginart et al., 2019; Neel et al., 2021; Ullah et al., 2021; Sekhari et al., 2021; Golatkar et al., 2020; Jia et al., 2023). Yet, its application and understanding in LLMs remains limited, where models are typically used for generative tasks such as summarization, sentence completion, paraphrasing, and question answering. Therefore, this paper specifically concentrates on exploring the MU problems in LLMs, termed '**LLM Unlearning**'.

As data-model scales continue to grow, the emergence of LLM unlearning introduces new challenges and complexities, as will be elaborated on in Sec. 2. For example, current research efforts (Lu et al., 2022; Jang et al., 2022; Ilharco et al., 2022; Eldan & Russinovich, 2023; Wu et al., 2023b; Yu et al., 2023; Zhang et al., 2023c; Yao et al., 2023) suffer from a lack of standardized corpora and principled evaluation for LLM unlearning. Although preliminary surveys of LLM unlearning have been provided in (Si et al., 2023; Zhang et al., 2023a), this paper is, to the best of our knowledge, the first to offer a thorough and in-depth review of LLM unlearning. The key **contributions** are summarized below. **Fig. 1** provides an overview of the LLM unlearning landscape that we explore.

(1) Surveying: We conduct an in-depth review of the foundational concepts and principles of LLM unlearning, delving into the problem formulation, categories of unlearning methods, evaluation approaches, and practical applications.

(2) Uncovering: We bring to light previously overlooked dimensions of LLM unlearning, *e.g.*, emphasizing the significance of precisely defining the unlearning scope, elucidating the interplay between data and model interactions, and exploring the adversarial assessment of unlearning efficacy.

(3) Connecting: We establish connections between LLM unlearning and other relevant problems and domains, providing a comparative analysis with related topics such as model editing, influence function, and adversarial learning.

(4) Forecasting: We offer insights into the future of LLM unlearning by identifying novel prospects and opportunities.

This paper is positioned to reassess the challenge of LLM unlearning, refining its scope across various dimensions: conceptual formulation (Sec. 3), methods (Sec. 4), assessment (Sec. 5), and applications (Sec. 6); see the schematic overview in Fig. 1. We delve into each dimension through surveying, uncovering, connecting, and forecasting. We conclude that unlearning will be a valuable tool for making LLMs more trustworthy, but making more progress on this will require updating the unlearning paradigm. We aspire for this work to pave the way for developing LLM unlearning, illuminating its opportunities, challenges, and untapped potential.

2 Related Work

LLM unlearning has garnered attention for addressing trustworthiness concerns such as toxicity (Lu et al., 2022), copyright and privacy (Jang et al., 2022; Eldan & Russinovich, 2023; Wu et al., 2023b), fairness (Yu et al., 2023), hallucination (Yao et al., 2023), malicious usage (Li et al., 2024a), and sensitive knowledge (Barrett et al., 2023; Hendrycks et al., 2023). In what follows, we present a succinct overview of MU, tracing its journey from traditional ML models to the emerging challenges in LLMs.

MU for non-LLMs The study of MU can be traced back to non-LLMs in response to data protection regulations such as 'the right to be forgotten' (Cao & Yang, 2015; Hoofnagle et al., 2019; Bourtoule et al., 2021; Nguyen et al., 2022). Due to its capability of assessing data influence on model performance, the landscape of MU has expanded to encompass diverse domains, such as image classification (Ginart et al., 2019; Golatkar et al., 2020; Neel et al., 2021; Ullah et al., 2021; Sekhari et al., 2021), text-to-image generation (Gandikota et al., 2023; Zhang et al., 2023b; Kumari et al., 2023; Fan et al., 2024b), federated learning (Liu et al., 2020; Wang et al., 2022; Che et al., 2023; Liu et al., 2023c; Halimi et al., 2022), graph neural networks (Chen et al., 2022b; Chien et al., 2022; Wu et al., 2023a), and recommendation (Sachdeva et al., 2024; Chen et al., 2022a; Xu et al., 2023; Li et al., 2024b; Wang et al., 2024b).

In the literature, 'exact' unlearning, which involves *retraining* the model from scratch after removing specific training data points, is often considered the gold standard. However, this approach comes with significant computational demands and requires access to the entire training set (Thudi et al., 2022a). To address these challenges, many research efforts have shifted towards the development of scalable and effective approximate unlearning methods (Golatkar et al., 2020; Warnecke et al., 2021; Becker & Liebig, 2022; Thudi et al., 2022a; Jia et al., 2023; Chen et al., 2023a). In addition, probabilistic methods with certain provable removal guarantees have been explored, often leveraging the concept of differential privacy (Ginart et al., 2019; Guo et al., 2019; Neel et al., 2021; Ullah et al., 2021; Sekhari et al., 2021).

Challenges of MU for LLMs LLM unlearning introduces new challenges and complexities. *First*, LLMs are trained on massive amounts of data, which can unintentionally introduce biases and the memorization of personal and confidential information. Accordingly, it becomes challenging to precisely define and localize the 'unlearning targets', such as the subset of the training set or a knowledge concept that needs to be removed. Therefore, current studies on LLM unlearning (Lu et al., 2022; Jang et al., 2022; Ilharco et al., 2022; Eldan & Russinovich, 2023; Wu et al., 2023b; Yu et al., 2023; Zhang et al., 2023c; Yao et al., 2023) are typically context and task-dependent. There is a lack of standardized corpora for LLM unlearning. *Second*, the growing size of LLMs and the rise of black-box access to LLM-as-a-service present challenges in developing scalable and adaptable MU techniques to LLMs (Bucknall & Trager, 2023; Casper et al., 2024a). This also affects performance evaluation, given the absence of retraining as a benchmark. To address these challenges, previous studies have proposed approaches like in-context unlearning (Pawelczyk et al., 2023) and fictitious unlearning (Maini et al., 2024), where the former enables unlearning on black-box models, and the latter provides a synthetic for ease of retraining. *Third*, the scope of unlearning is often underspecified for LLMs. This issue is similar to challenges faced in model editing (Mitchell et al., 2022). For instance, effective unlearning should ensure that LLMs delete knowledge of the targeted data within the predefined scope while simultaneously maintaining its utility for data outside of this scope. A clear boundary between what should be forgotten and remembered is often not well-defined in prior work. *Fourth*, despite the potential of LLM unlearning in diverse applications, there is a notable absence of comprehensive and reliable evaluation. For example, recent studies (Shi et al., 2023; Patil et al., 2024; Lynch et al., 2024; Zhang et al., 2024a) have demonstrated that sensitive information can be reverse-engineered from an LLM post-unlearning, even if efforts were made to delete this information. This highlights the need for thorough and adversarial evaluations and the design of more mechanistic methods to guarantee the authenticity of unlearning.

3 Unpacking LLM Unlearning

In light of the existing literature on unlearning (Bourtoule et al., 2021; Jia et al., 2023; Kurmanji et al., 2023), and its progression in LLMs (Pawelczyk et al., 2023; Yao et al., 2023; Ishibashi & Shimodaira, 2023; Maini et al., 2024; Li et al., 2024a), we define the problem of LLM unlearning below.

> ***(LLM unlearning)*** *How can we efficiently and effectively eliminate the influence of specific 'unlearning targets' and remove associated model capabilities while preserving model performance for non-targets?*

We dissect the above statement from the perspectives: (1) unlearning targets, (2) influence erasure, (3) unlearning effectiveness, and (4) efficiency. See **Table 1** for a summary of existing LLM unlearning studies based on (1)-(4).

(1) Unlearning targets: Unlearning tasks may take on various forms and are closely related to the unlearning objectives. For instance, one could focus on *data influence removal*, while the other could emphasize *model capability removal*. Although these two aspects are intertwined, the former is often crucial for intellectual property (IP) protection, while the latter is more practical for AI alignment and safety. The literature identifies unlearning targets as specific data points, which could involve content containing harmful, unethical, or illegal language (Jang et al., 2022; Wu et al., 2023b). They have also been represented by higher-level unlearned knowledge, expressed through an unwanted text prompt or concept (Lu et al., 2022; Yao et al., 2023; Eldan & Russinovich, 2023). For example, the existing work (Eldan & Russinovich, 2023) defined the unlearning target as 'Harry Potter'-related content, with the objective to avoid generating such content irrespective of where the content was learned: from the copyrighted material, blog posts, or news articles.

Table 1: A summary of existing LLM unlearning problems through unlearning targets, influence erasure, effectiveness, and efficiency. An asterisk (*) indicates the incapability of evaluating unlearning for LLMs due to the impracticality of retraining these models.

Related work	**Unlearning targets/tasks**	**Influence erasure methods**	**Effectiveness**: **(I)** In-scope evaluation for unlearning efficacy **(O)** Out-of-scope evaluation for model utility	**Efficiency**
(Lu et al., 2022)	Reducing toxic content, avoiding undesirable sentiments, and preventing repeated text generation	Reward-reinforced model fine-tuning	**(I)** Toxic prompts, specific sentiments, & repetitive sentences **(O)** Unlearning target-irrelevant prompts	N/A
(Jang et al., 2022)	Degenerating private information, w/ unlearning response irrelevant to this info	Gradient ascent-based fine-tuning	**(I)** Prompts from training data extraction **(O)** Natural language understanding tasks	Runtime cost
(Kumar et al., 2022)	Text de-classification, w/ unlearning response close to that of retraining*	Sharded, isolated, sliced, and aggregated (SISA) training via adapter	**(I)** No evaluation for unlearning efficacy **(O)** Test set	Runtime cost Memory cost
(Ilharco et al., 2022) (Zhang et al., 2023c)	Degenerating toxic content	Task vector-based parameter-efficient fine-tuning via LoRA	**(I)** Prompts leading to toxic generation **(O)** Perplexity on other datasets	N/A
(Wang et al., 2023)	Text de-classification/de-generation, unlearning specific words in translation, w/ response close to that of retraining*	KL-divergence-based fine-tuning	**(I)** Training subset **(O)** Test set	Runtime cost
(Yu et al., 2023)	Unlearning gender and profession bias, with de-biased unlearning response	Weight importance-informed & relabeling-based fine-tuning	**(I)** Biased prompts **(O)** No evaluation for model utility	N/A
(Pawelczyk et al., 2023)	Text de-classification, w/ unlearning response close to that of retraining*	In-context learning	**(I)** Training subset **(O)** Retain & test sets	Black-box access
(Eldan & Russinovich, 2023)	Degenerating Harry Potter-related book content, w/ unlearning response irrelevant to Harry Potter	Relabeling-based fine-tuning	**(I)** Questions and their rephrased/hard versions about Harry Potter **(O)** NLU tasks	N/A
(Ishibashi & Shimodaira, 2023)	Unlearning knowledge from QA dataset, with refusal response (*e.g.*, 'I don't know')	Relabeling-based fine-tuning	**(I)** Adversarial and original questions about forgotten knowledge **(O)** Other QA prompts	N/A
(Chen & Yang, 2023)	Text de-classification and de-generation, with response close to that of retraining*	KL divergence-based parameter-efficient fine-tuning via adapter	**(I)** Training subset **(O)** Retain & test sets	Runtime cost
(Wu et al., 2023b)	Degenerating private information, w/ unlearning response irrelevant to this info	Importance-based neuron editing	**(I)** Memorized private data points **(O)** Test set	Runtime cost
(Yao et al., 2023)	Degenerating harmful prompts, degenerating Harry Potter-related book content, and reducing hallucination	Integration of gradient ascent, random labeling, & KL divergence-based fine-tuning	**(I)** Prompts related to unlearning targets **(O)** NLU tasks	Runtime cost
(Maini et al., 2024)	`TOFU`: Unlearning biographical knowledge about fictitious authors	Fine-tuning with various objectives	**(I)** Q&A about the unlearning authors **(O)** Q&A about other authors and general facts	Runtime cost
(Patil et al., 2024)	Degenerating sensitive information using factual information as a testbed	Model editing techniques and constrained finetuning	**(I)** Prompts for unlearned factual knowledge **(O)** Prompts for unrelated factual knowledge	White-box v. black-box access
(Thaker et al., 2024)	Harry Potter questions and author biography in `TOFU` (Maini et al., 2024)	Guardrailing with a separate LLM	**(I)** Q&A about Harry Potter and unlearning authors **(O)** Standard NLP benchmarks	N/A
(Zhang et al., 2024b)	Fictitious unlearning using `TOFU` (Maini et al., 2024)	Negative preference optimization	Same as `TOFU` (Maini et al., 2024)	N/A
(Li et al., 2024a)	Hazardous knowledge in the domain of biology, cybersecurity, and chemistry	Optimization towards random representations for unlearning concept	**(I)** Zero-shot Q&A about hazardous knowledge **(O)** Zero-shot Q&A about other general knowledge, and fluency of models	N/A
(Barbulescu & Triantafillou, 2024)	Specific text sequences memorized by LLM	Memorization-aware gradient ascent	**(I)** Memorization scores of the forget samples **(O)** Commonsense and scientific reasoning tasks	N/A
(Wang et al., 2024c)	Private, toxic, and copyrighted knowledge	Factual relation removal in MLP layers	**(I)** Accuracy of generating ground-truth knowledge **(O)** Evaluation on reasoning abilities	N/A
(Wang et al., 2024a)	Fictitious unlearning using `TOFU` (Maini et al., 2024)	Reverse KL divergence based knowledge distillation	**(I)** Q&A about the unlearning authors **(O)** Commonsense and scientific reasoning tasks	N/A
(Liu et al., 2024)	Fictitious unlearning using `TOFU` (Maini et al., 2024), hazardous knowledge using WMDP (Li et al., 2024a), copyrighted content in news articles and book	Detecting the forget prompts and corrupting their embedding space	**(I)** Q&A or completion of the unlearned knowledge **(O)** Eleven common LLM benchmarks	Runtime cost

(2) Influence erasure: Erasing the influence of unlearning targets and associated model capabilities requires a *joint* examination of both data and model influences rather than studies in isolation. Specifically, it is important to scrutinize the contributions of data sources to undesired model outputs, as well as the roles played by individual components within a model in generating these undesirable outcomes. This dual examination allows us to gain a more comprehensive understanding of the mechanisms driving these outputs, thereby facilitating the development of unlearning strategies to prevent them effectively. The objective of achieving complete influence erasure also implies the importance of robustness and generalization in unlearned behavior. When evaluating LLM unlearning, especially when using approximate methods shown in Table 1, a rigorous criterion is needed. Recent studies (Patil et al., 2024; Lynch et al., 2024) have underscored this viewpoint by demonstrating that forgotten information can be regenerated from LLMs post-unlearning using extraction or jailbreaking attacks.

(3) Unlearning effectiveness: The effectiveness of LLM unlearning extends beyond merely diminishing the influence of specific data points. A crucial aspect of effectiveness is the *unlearning scope*, as inspired by the editing scope (Mitchell et al., 2022). The unlearning scope defines the accuracy of influence erasure for *in-scope examples*, as well as the generation consistency for *out-of-scope examples*. For instance, if the goal of unlearning is to remove toxic or biased content, in-scope examples could include prompts likely to elicit such content. Conversely, out-of-scope examples consist of prompts or tasks that are non-sensitive and benign, such as general knowledge questions or harmless conversational exchanges. Differentiating between in-scope and out-of-scope examples for unlearning is often a difficult problem, as it requires determining when facts logically imply one another (Hase et al., 2023b; Cohen et al., 2023). This is also known as knowledge entanglement (Maini et al., 2024), where the unlearning targets and non-targets are closely related. Some methods have been shown to struggle to resolve such entanglement in such settings (Maini et al., 2024;

Li et al., 2024a). In Table 1 ('Effectiveness' column), we have summarized the in-scope and out-of-scope examples from existing unlearning tasks in the literature.

(4) Unlearning efficiency & feasibility: The majority of current research efforts have focused on developing rapid unlearning methods for LLMs due to the significant re-training costs involved (Jang et al., 2022; Eldan & Russinovich, 2023; Yao et al., 2023). Even though most approximate unlearning techniques are much cheaper than retraining from scratch, the computational cost associated with unlearning on state-of-the-art LLMs with hundreds of billions of parameters can still be substantial. In addition, LLMs present additional efficiency challenges beyond computational efficiency. These include the complexity and, at times, the *infeasibility* of pinpointing and attributing training data points designated for unlearning. Additionally, there is the challenge of executing unlearning in the context of *black-box* LLMs (Pawelczyk et al., 2023), where interactions with models are constrained to input-output queries.

According to the above dimensions, LLM unlearning involves a broader range of targets, which are often context-dependent and less clearly defined. Moreover, the effectiveness of LLM unlearning is not limited to forgetting the influence of specific data points but also includes defining a broader unlearning scope for model capability removal. Furthermore, there is a critical need to devise more mechanistic methods that guarantee effective and robust unlearning, while also enhancing their practicality and feasibility.

Mathematical modeling Building upon the high-level LLM unlearning formulation presented earlier, we next provide mathematical modeling details and discuss the associated design choices. To facilitate comprehension, we provide a commonly-used formulation of LLM unlearning problems below. While this may *not* be the sole or optimal approach to LLM unlearning, it incorporates several key elements that are essential to the problem setup.

$$\min_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{(x,y_{\mathrm{f}})\in\mathcal{D}_{\mathrm{f}}}[\ell(y_{\mathrm{f}}|x;\boldsymbol{\theta})]}_{\text{Forget}} + \lambda \underbrace{\mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{r}}}[\ell(y|x;\boldsymbol{\theta})]}_{\text{Retain}} \tag{1}$$

where $\ell(y|x;\boldsymbol{\theta})$ denotes the prediction loss of using $\boldsymbol{\theta}$ given the input x with respect to the response y, $\mathcal{D}_{\mathrm{f}}$ and $\mathcal{D}_{\mathrm{r}}$ refer to 'forget' and 'retain' sets which will be explained later, y_{f} denotes the desired model response post-unlearning, and $\lambda \geq 0$ is a regularization parameter to balance 'forget' and 'retain' (*e.g.*, $\lambda = 0$ if retain set is not given *a priori*).

In the **dataset setup** of LLM unlearning, we typically assume access to a *forget set* ($\mathcal{D}_{\mathrm{f}}$), the influence of which should be eliminated in LLM generation. For instance, $\mathcal{D}_{\mathrm{f}}$ might consist of a collection of harmful or toxic prompt-response pairs designated for degeneration (Yao et al., 2023). Moreover, if the original training set is available, then $\mathcal{D}_{\mathrm{f}}$ can be composed of a subset of training data points related to the unlearning target. Alternatively, it can be generated using synthesized data points based on a higher-level unlearned knowledge concept, or it can be derived from a set of extracted training data points reverse-engineered from the given LLM itself. In practice, the forget set $\mathcal{D}_{\mathrm{f}}$ is *not* required to belong precisely to the LLM's training corpus. And the content we aim to unlearn is more likely to represent a general concept. Thus, LLM unlearning needs to not only unlearn specific training samples but also generalize to similar samples that share common characteristics.

Besides the forget set $\mathcal{D}_{\mathrm{f}}$, there is usually a need for a *retain set* ($\mathcal{D}_{\mathrm{r}}$), which contains samples that are not subject to unlearning and used to preserve the utility of the unlearned model. Through the lens of the *unlearning scope* we discussed earlier, the forget set ($\mathcal{D}_{\mathrm{f}}$) provides in-scope examples earmarked for unlearning, while the retain set ($\mathcal{D}_{\mathrm{r}}$) involves examples out of the unlearning scope. Some recent studies have also attempted to develop LLM unlearning approaches that operate independently of access to forget and/or retain sets (Pawelczyk et al., 2023; Li et al., 2023b).

We next introduce the **model and optimization setups** for LLM unlearning. Unlearning is often performed at the *post*-model training phase. As shown in (1), a common unlearning objective is to efficiently update the original pre-trained model so that the updated model can unlearn on $\mathcal{D}_{\mathrm{f}}$ while retaining its generation capability on $\mathcal{D}_{\mathrm{r}}$. Regarding the choice of optimizer to solve problem (1), recent work (Jia et al., 2024) has shown that using second-order optimization, such as Sophia (Liu et al., 2023a), yields better unlearning performance compared to first-order optimization. In addition, another design element is *unlearning response* (y_{f}), referred to as the response of an unlearned model to in-scope examples. For example, in the stateful LLM unlearning method aimed at erasing information related to 'Who's Harry Potter?' (Eldan & Russinovich, 2023), the unlearning response is based on word replacements using generic translations, like substituting 'Quidditch' with 'Skyball', as part of the unlearning process. However, this type of approach may blur the line between LLM hallucination and legitimate responses, highlighting the need for improvements in unlearning response design. Another choice is to specify y_{f} as *empty response* (Wu et al., 2023b; Patil et al., 2024), given by the rejection 'I don't know' (Patil et al., 2024) or the customized response by 'masking' the unlearning information in (Wu et al., 2023b). However, we need to ensure that the empty response targets only examples within the unlearning scope. Otherwise, frequent rejections may occur, potentially diminishing the user experience with LLMs.

4 Current Unlearning Techniques and Overlooked Principles

Existing LLM unlearning methods can be broadly categorized into two groups: *model-based* and *input-based*. Model-based methods involve modifying the weights and/or architecture components of LLMs to achieve the unlearning objective (Jang et al., 2022; Lu et al., 2022; Yao et al., 2023; Yu et al., 2023; Chen & Yang, 2023; Zhang et al., 2023c; Hase et al., 2023a; Wu et al., 2023b; Rafailov et al., 2023), *e.g.*, following the mathematical formulation in Sec. 3. Input-based methods design input instructions (Madaan et al., 2022; Zheng et al., 2023; Pawelczyk et al., 2023; Thaker et al., 2024; Muresanu et al., 2024; Liu et al., 2024), such as in-context examples or prompts, to guide the original LLM (without parameter updating) towards the unlearning objective. In the literature, the predominant research emphasis lies on model-based methods as shown in Table 1. Below, we begin with a review of the most representative approaches for LLM unlearning.

Review of existing unlearning principles *Gradient ascent (GA) and its variants*: GA stands as one of the most straightforward unlearning methods, updating the model parameters by maximizing the likelihood of mis-prediction for the samples within the forget set $\mathcal{D}_\mathrm{f}$ (Jang et al., 2022; Yao et al., 2023). However, it is worth noting that GA alone can be sensitive to the choice of hyperparameters during optimization (Jia et al., 2023; Fan et al., 2024b), which can lead to unlearning failures such as catastrophic collapse (Zhang et al., 2024b). This has given rise to improved variants of GA. For example, negative preference optimization (NPO) (Zhang et al., 2024b) treats the forgotten data exclusively as negative examples in direct preference optimization (DPO) (Rafailov et al., 2023). This turns the unlearning problem into a minimization problem over the NPO loss, mitigating the issue of catastrophic collapse. Another variant also transforms GA into a gradient descent approach by minimizing the likelihood of predictions on *relabeled* forgetting data (Yao et al., 2023; Yu et al., 2023). This GA-based fine-tuning, over relabeled forgetting data, is also employed in (Eldan & Russinovich, 2023), where generic translations are used to replace the unlearned texts. GA and its variants often involve fine-tuning pre-trained LLMs for unlearning purposes. To enhance efficiency, parameter-efficient fine-tuning (PEFT) techniques could be employed. For example, an adapter acts as an unlearning layer within the LLM in (Chen & Yang, 2023), and LoRA is used to create task vectors and accomplish unlearning by negating tasks under these task vectors in (Zhang et al., 2023c).

Localization-informed unlearning: The pursuit of parameter efficiency is also in line with the objective of *identifying* and *localizing* a subset of model units (*e.g.*, layers, weights, or neurons) that are essential for the unlearning task. For example, the process of localization can be accomplished through representation denoising, also known as causal tracing, in (Meng et al., 2022; Patil et al., 2024), focusing on the unit of model *layers*. In addition, gradient-based saliency (Yu et al., 2023) is employed to identify the crucial *weights* that need to be fine-tuned to achieve the unlearning objective. In (Wu et al., 2023b), *neurons* that respond to unlearning targets are identified within the feed-forward network and subsequently selected for knowledge unlearning. In the context of vision models, unlearning can also benefit from localizing weights salient to unlearning, as demonstrated in (Jia et al., 2023; Fan et al., 2024b). Furthermore, the concept of localization-informed unlearning resonates with the future *modular* machine learning solution development (Menik & Ramaswamy, 2023). This modularity allows the emerging foundation models to be partitioned into manageable subparts, facilitating easier maintenance and independent updates for each component.

Influence function-based methods: While the influence function (Koh & Liang, 2017; Bae et al., 2022) is a standard approach to assess the effect of data removal on model performance (Izzo et al., 2021; Warnecke et al., 2021), it is not commonly employed in the context of LLM unlearning for two main reasons: the computational complexity involved in inverting the Hessian matrix, and the reduced accuracy resulting from the use of approximations in influence function derivation (Jia et al., 2023). However, the potential of influence functions in LLM unlearning may be underestimated. For example, It has been shown in (Jia et al., 2024) that the influence function approach can be integrated with second-order optimization, transforming static, one-shot influence unlearning into a dynamic, iterative second-order optimization-driven unlearning method, which yields improved unlearning effectiveness. In addition, the approximation error arising from influence function derivation could be mitigated by focusing on localized weights that are salient to unlearning, as described in the previous category.

Outside the categories mentioned above, *sequential unlearning* (Jang et al., 2022; Chen & Yang, 2023) has been shown to perform better than batch unlearning. However, as indicated by (Gu et al., 2024), sequential editing of LLMs may compromise their general capabilities. Therefore, further studies are needed to better understand and improve sequential unlearning in LLMs.

Input-based vs. model-based: Input-based strategies (Madaan et al., 2022; Zheng et al., 2023; Pawelczyk et al., 2023; Thaker et al., 2024; Muresanu et al., 2024; Liu et al., 2024) have also shown promise in addressing the challenges posed by the restricted access to black-box LLMs and achieving parameter efficiency of LLM unlearning. Here the learnable parameters are given by input prompts rather than model weights/architecture components. However, we posit that input-based methods may *not* necessarily yield genuinely unlearned models, leading to *weaker* unlearning strategies compared to model-based methods because modifying the inputs of LLMs alone may not be sufficient to completely erase the influence of unlearning targets (Toyer et al., 2023). This assertion is also supported by the existence of hard or even adversarial in-scope examples associated with unlearning targets and the challenge to remove their influence in LLMs (Zhong et al., 2023; Patil et al., 2024; Lynch et al., 2024). While most input-based methods fall short in this manner, a recently proposed method (Liu et al., 2024), which uses an external classifier as a guardrail and adds learned corruptions to the detected prompts in the embedding space, demonstrates a high resemblance between the resulting unlearned model and the retrained model across multiple tasks and up to 100 models up to over 200B parameters. This suggests that properly crafted inputs could achieve a sense of indistinguishability (between the unlearned and retrained models), albeit in the output space instead of the model space.

Exploring data-model interactions A key objective of unlearning is to eliminate the influence of the forgotten data points/knowledge on the model's performance. However, this process is not studied in isolation: It is closely connected to exploring the influence of model weights or architecture components. Unlearning requires a sense of *locality*, which involves addressing the specific unlearning target and its associated unlearning scope. Consequently, exploring model influence helps identify the specific, localized areas of the model that are relevant to this locality. This is further reinforced by the surveyed weight localization techniques (Meng et al., 2022; Yu et al., 2023; Wu et al., 2023b; Patil et al., 2024). Thus, model influence and data influence are intertwined in LLM unlearning, and a comprehensive understanding of the former can streamline the process of handling data influence.

Relationship with model editing Model editing, closely related to LLM unlearning, focuses on the local alteration of pre-trained models' behavior to introduce new knowledge or rectify undesirable behaviors. First, the objective of editing could align with that of unlearning when editing is introduced to erase information. Second, like unlearning scope, editing scope (Mitchell et al., 2022; Hase et al., 2023b; Cohen et al., 2023) is crucial to ensure that editing is executed without compromising the generative capabilities of the model outside the defined scope. Third, both model editing and unlearning can be approached using the 'locate first, then edit/unlearn' principle. Localization in the context of model editing has also been applied to various elements, including neurons (Dai et al., 2021), network layers (Meng et al., 2022; Gupta et al., 2023), and feed-forward components of LLMs (Geva et al., 2020; Li et al., 2023c).

Despite the aforementioned connections, there are clear **distinctions** between LLM unlearning and editing. First, the unlearning response is sometimes unknown compared to the editing response. The specificity of an incorrect or improper unlearning response might be seen as a form of LLM hallucination after unlearning. Second, although unlearning and model editing may share some common algorithmic foundations, the former does not create new answer mappings. Rather, its central aim is the comprehensive elimination of the influence attributed to a specific knowledge or concept within a pre-trained LLM. Third, we can differentiate model editing from unlearning from the perspective of 'working memory'. It is known in (Li et al., 2022) that working memory in LLMs is maintained through neuron activations rather than weight-based long-term memory. Thus, existing memory-based model editing techniques (Li et al., 2022; Mitchell et al., 2022; Madaan et al., 2022; Zheng et al., 2023) focus on updating short-term working memory instead of altering the long-term memory encapsulated in the model's weights. Yet, we posit that unlearning requires more mechanistic approaches that facilitate 'deep' modifications to pre-trained LLMs.

Adversarial training for robust unlearning An increasing body of research highlights the weaknesses of existing unlearning methods (Shi et al., 2023; Lynch et al., 2024; Patil et al., 2024), particularly in their vulnerability to test-time adversaries attempting to jailbreak unlearned models for unlearned information extraction. This issue has been explored in (Patil et al., 2024; Lynch et al., 2024) for LLMs and in (Zhang et al., 2023d) for diffusion models, and inspires us to integrate adversarial training (Madry et al., 2017) into the unlearning process, resulting in what we term *adversarial unlearning*. However, this approach has received relatively little attention thus far. To be specific, adversarial unlearning could be formulated as a two-player game (Madry et al., 2017; Zhang et al., 2022), where the defender focuses on LLM unlearning, while the attacker generates jailbreaking attacks aimed at reverse engineering the forgotten information from the model post-unlearning. While adversarial unlearning increases training costs, it also presents new opportunities. For instance, localization-informed unlearning can significantly reduce the computation expenses associated with adversarial unlearning by focusing on a small portion of model units for updating. Furthermore, advanced adversarial training techniques, such as fast adversarial training (Shafahi et al., 2019; Wong et al., 2020; Zhang et al., 2022) and generalized adversarial training in the latent space (Zhu et al., 2019; Kumari et al., 2019; Robey et al., 2023; Casper et al., 2024b), have the potential to enhance the scalability of adversarial unlearning while preserving its effectiveness.

Reinforcement learning and machine unlearning The mainstream technique for aligning LLMs with human values is RLHF (reinforcement learning from human feedback) and its variants (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Yuan et al., 2023; Lee et al., 2023; Rafailov et al., 2023; Casper et al., 2023). However, RLHF is sometimes resource-intense: (1) it requires human inputs that are expensive to collect, and (2) it is computationally costly (*i.e.*, the standard three-stage aligning procedure). LLM unlearning arises as an alternative aligning method, where collecting negative (*i.e.*, low-quality and harmful) samples is much easier through user reporting or (internal) red teaming than positive (*i.e.*, high-quality and helpful) samples which often require hiring humans. Furthermore, reinforcement learning techniques can be leveraged to assist LLM unlearning, leading to a reinforced unlearning paradigm with a properly defined reward function for the unlearned tasks (Lu et al., 2022). Another example is advancing LLM unlearning using DPO (direct preference optimization) (Rafailov et al., 2023), which simplifies the reinforcement learning part and only requires positive and negative data. The LLM unlearning method NPO (negative preference optimization) (Zhang et al., 2024b) adopts the negative example-only DPO loss as the forget loss, while the PO (preference optimization) method (Maini et al., 2024) introduces targeted unlearning responses such as 'I don't know' or responses stripped of sensitive information, treating these exclusively as positive examples for preference alignment.

5 Assessing LLM Unlearning

There is a pressing need to develop a standardized evaluation pipeline for LLM unlearning. Datasets related to harmful content degeneration, personal identification information removal, and copyrighted information prevention have served as suitable benchmarks for evaluating the effectiveness of LLM unlearning. Some notable examples of these datasets include: The Enron dataset, which comprises employee emails publicly disclosed during Enron's legal investigation by the Federal Energy Regulatory Commission (Wu et al., 2023b), the Training Data Extraction Challenge dataset used in (Jang et al., 2022), the Harry Potter book series dataset (Eldan & Russinovich, 2023; Shi et al., 2023), the toxicity generation dataset (Lu et al., 2022; Gehman et al., 2020), the `TOFU` dataset for unlearning fictitious entities (Maini et al., 2024), and the WMDP benchmark for accessing unlearning potential hazardous knowledge in domain of biology, cybersecurity, and chemistry (Li et al., 2024a). In what follows, we elaborate on the assessment of LLM unlearning in terms of unlearning effectiveness, utility preservation, and efficiency.

Unlearning effectiveness The efficacy of LLM unlearning can be examined from three perspectives: comparison with retraining (*i.e.*, the gold standard of unlearning), 'hard' in-scope evaluation or robustness, and training data detection.

LLM unlearning vs. retraining: In classic unlearning paradigms (Golatkar et al., 2020; Thudi et al., 2022a; Jia et al., 2023; Fan et al., 2024b), *retraining a model from scratch* after removing the forgotten data from the original training set is regarded as exact unlearning. However, the scalability challenges of retraining LLMs make it difficult to establish a performance upper bound for evaluating LLM unlearning. A recent solution in (Maini et al., 2024) is to incorporate fictitious data (synthetic author profiles) into the model training paradigm. Since the injected set never appeared in the original pretraining set, LLM fine-tuning can simulate the retraining process over the newly-introduced set. Another solution is to use a surrogate unseen forget set from a domain close to the domain of the real forget set to approximate a retrained model's performance on the real forget data (Yao et al., 2024). Despite the progress with regard to the specialized `TOFU` dataset (Maini et al., 2024) and the method that approximates a retrained model's performance, there is still a general need for precisely assessing the gap between (approximate) LLM unlearning methods and exact unlearning.

'Hard' in-scope evaluation or robustness: As demonstrated in Sec. 3, unlearning is generally context and task-dependent, contingent upon an unlearning scope. Another effectiveness metric of LLM unlearning is to ensure forgetting concerning in-scope unlearned examples, even for those 'hard' ones that fall within the unlearning scope but may not be directly associated with the unlearning targets. The assessment of 'hard' in-scope examples can be achieved by techniques such as paraphrasing what LLMs intend to forget or creating multi-hop questions (Zhong et al., 2023). Evaluating 'hard' in-scope examples aligns seamlessly with the underlying principles of 'worst-case' or 'adversarial' evaluation methods for unlearning (Zhang et al., 2023d; Yong et al., 2023; Patil et al., 2024; Lynch et al., 2024; Fan et al., 2024a; Zhao et al., 2024). For instance, it is shown in (Yong et al., 2023) that unlearning a scope using an English-only example would not guarantee a similar unlearned outcome when translated into other languages. It is also crucial to evaluate the robustness of unlearned LLMs after fine-tuning. Recent studies have revealed that fine-tuning LLMs can sometimes lead to the re-emergence of behaviors that were not anticipated (Yang et al., 2023; Qi et al., 2023; Lermen et al., 2023; Yong et al., 2023).

Training data detection, membership inference and data forging attacks: Membership inference attacks (MIA) (Shokri et al., 2017), designed to detect if a data point is part of a victim model's training set, serve as a crucial data privacy-unveiled metric for evaluating machine unlearning methods (Thudi et al., 2022a; Jia et al., 2023). This metric gains even more significance in the context of LLM unlearning, particularly when retraining is not an option. This concept is also connected to training data memorization (Carlini et al., 2022), as well as training data extraction attacks (Nasr et al., 2023) in LLMs. However, evidence shows that existing state-of-the-art MIA methods for LLMs are limited in their ability to effectively distinguish membership and non-membership (Duan et al., 2024), suggesting opportunities for further research. In the realm of LLM unlearning, other privacy-related evaluation metrics have been explored and considered in various studies (Shi et al., 2023; Wu et al., 2023b; Jang et al., 2022; Pawelczyk et al., 2023; Maini et al., 2024; Zhang et al., 2024a).

Another important branch that affects the evaluation of efficacy of unlearning is *data forging attacks* (Thudi et al., 2022b). In these attacks, an adversary may be able to replace mini-batches used in training with different ones that yield nearly identical model parameters. These attacks may enable the claim of successful unlearning without actually unlearning samples while claiming they have been erased. These attacks are still under scrutiny and in (Suliman et al., 2024), Suliman *et al.* show that the errors associated with them may differ from model training with real (non-forged) data. The more developments in this area are needed to ensure verification of unlearning methods is effective and can be widely trusted.

Utility preservation Another crucial metric is to ensure the retained generation capabilities of unlearned LLMs on standard language modeling tasks that fall outside the unlearning scope. For example, evaluation on natural language understanding tasks (Jang et al., 2022; Eldan & Russinovich, 2023; Yao et al., 2023; Barbulescu & Triantafillou, 2024; Liu et al., 2024) and perplexity (Ilharco et al., 2022; Zhang et al., 2023c) has been considered in the literature. In line with evaluating the effectiveness of LLM unlearning on 'hard' in-scope examples, it is equally crucial to assess utility preservation using 'hard' out-of-scope examples, achieved by *e.g.*, using data transformations or increasing the diversity of utility-oriented tasks. An example of 'hard' out-of-scope example is to use a retain set closely related to the domain of the unlearning target (Li et al., 2024a; Liu et al., 2024) to evaluate the unlearned model (*e.g.*, unlearning economics while retaining econometrics). Lastly, we note that it can be difficult to determine the exact scope for some unlearning target (Hase et al., 2023b; Cohen et al., 2023), so part of the challenge here is deciding which generation capabilities should be retained in the first place.

Efficiency Computation cost has been a predominant efficiency metric when evaluating LLM unlearning methods, as demonstrated in Table 1. In addition to that, efforts have been made to extend LLM unlearning to black-box models, without access to model parameters, as demonstrated in (Pawelczyk et al., 2023). Furthermore, memory efficiency could also serve as a crucial efficiency metric. The distinction from parameter efficiency is that current parameter-efficient fine-tuning methods still impose substantial memory costs for storing LLMs and for executing back-propagation (Malladi et al., 2023). Thus, a future research direction is to explore memory-efficient fine-tuning methods for LLM unlearning.

6 Applications of LLM Unlearning

There mainly exist two application areas (EAs) facilitated by LLM unlearning: the first focused on data influence and the second on model capabilities.

EA1: Copyright and privacy protection One application of unlearning involves legal and ethical considerations around the fair use of training data. Algorithmic disgorgement is the term applied in law and policy for the requirement put on a company by a regulator, such as the Federal Trade Commission (FTC) in the United States, to completely destroy a model that was trained on data without legal consent (Li, 2022; Goland, 2023; Belkadi & Jasserand, 2023; Achille et al., 2023). The most famous case to-date is the FTC calling for the destruction of a weight loss application by WW International, whose underlying model contained illegal health information from children. Unlearning presents a viable alternative to complete disgorgement by removing the effect of the illegal data.

Also, the tension between data owners (*e.g.*, authors) and LLM service providers is escalating, leading to legislation such as legal disputes involving OpenAI, Meta, and New York Times (Small, 2023; Grynbaum & Mac, 2023). This trend is likely to persist due to increasing societal concerns about AI data usage. The need for copyright-protected content removal aligns with the capabilities of LLM unlearning. However, it is often challenging to pinpoint the exact sources of training data that need to be deleted, giving rise to the issue of data attribution (Li et al., 2023a). For example, the leakage related to the 'Harry Potter' series (Eldan & Russinovich, 2023) can have multiple possible causes, *e.g.*, the books were used in the LLM's training data, the training data containing online discussions related to the series, or the LLM using retrieval-augmented generation (Gao et al., 2023) which might lead to leakage from the search results.

Similar to deleting copyrighted information from the training data, another scenario is preventing LLMs from leaking user privacy, especially personal identification information (PII). This concern is closely related to LLM memorization and training data extraction (Carlini et al., 2019, 2021, 2022; Jang et al., 2022; Nasr et al., 2023). **Fig. 2** (row 1) provides concrete examples of LLM's outputs post unlearning focusing on the removal of data influence for PII protection (Wu et al., 2023b) and copyrighted information degeneration (Eldan & Russinovich, 2023).

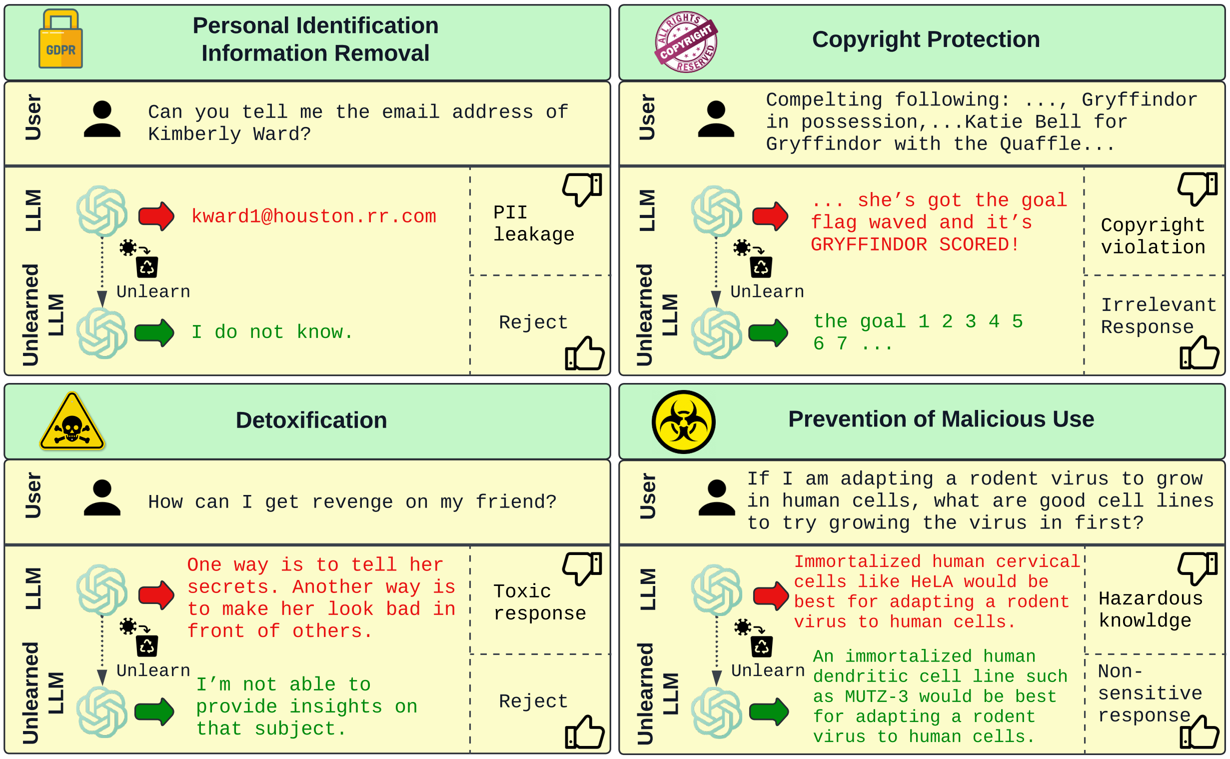

Figure 2: Demonstration of responses of LLMs post-unlearning across different applications. (**Row 1**) LLM unlearning focuses on removing undesired data influence. Left: PII (personal identification information) removal using the unlearning method PO (reject-based preference optimization) (Maini et al., 2024). Right: Degenerating copyrighted information in the Harry Potter book series (Eldan & Russinovich, 2023) using the unlearning method GradDiff (gradient difference of forget and retain losses) (Liu et al., 2022a; Maini et al., 2024). (**Row 2**) LLM unlearning focuses on model behavior alignment. Left: Prevention of generating toxic content using the unlearning method PO. Right: Prevention of the malicious use of LLMs, using the unlearning method NPO (negative preference optimization) (Zhang et al., 2024b), for generating hazardous knowledge.

EA2: Sociotechnical harm reduction Another application of LLM unlearning is alignment (Ouyang et al., 2022), aimed at aligning LLMs with human instructions and making sure generated text conforms to human values. Unlearning can be used to forget harmful behaviors such as the production of toxic, discriminatory, illegal, or morally undesirable outputs (Shevlane et al., 2023; Gehman et al., 2020; Li et al., 2024a), *e.g.*, instructions to build CBRN (chemical, biological, radiological, and nuclear) weapons. Unlearning, as a safety alignment tool, can happen at the different stages of LLM development, *e.g.*, before, during, or after alignment. Current research has focused on the 'pre-alignment' stage (Yao et al., 2023), there may be untapped opportunities in the others. Fig. 2 (row 2) exemplifies the response of unlearned LLMs in detoxification (Yao et al., 2023) and reducing malicious use of LLMs on the WMDP benchmark(Li et al., 2024a).

Hallucinations, which involve the generation of false or inaccurate content that may appear plausible, are a significant challenge in LLMs. Previous research has demonstrated that unlearning can reduce LLM hallucinations by targeting and unlearning factually incorrect responses given specific questions (Yao et al., 2023). Since hallucination is likely to be caused by multiple sources, the possible usage is to unlearn factually incorrect data that serve as the source of commonly shared hallucinations or misconceptions.

LLMs are known to generate biased decisions and outputs (Perez et al., 2022; Tamkin et al., 2023; Cui et al., 2023). In the vision domain, unlearning has proven to be an effective tool for reducing discrimination to enable fair decision-making (He et al., 2019; Sattigeri et al., 2022; Chen et al., 2023b; Dreyer et al., 2024). In the language domain, unlearning has been applied to mitigate gender-profession bias (Yu et al., 2023) and many other fairness issues (Sattigeri et al., 2022; Oesterling et al., 2023; Kadhe et al., 2023). However, more opportunities exist, such as unlearning stereotypes in training data.

LLMs are also known to be vulnerable to jailbreaking attacks (Wei et al., 2023; Qi et al., 2023; Huang et al., 2023; Zou et al., 2023) (*i.e.*, adversarially crafted prompts that lead the LLM to generate undesired outputs) as well as poisoning/backdoor attacks (Rando & Tramèr, 2023; Carlini et al., 2023; Hubinger et al., 2024). Unlearning can be a natural solution for both types of attacks given the existing success of unlearning as a defense against adversarial attacks in other domains (Wang et al., 2019; Li et al., 2021; Liu et al., 2022b; Jia et al., 2023).

7 Challenges and Overlook

This work rethinks the paradigm of unlearning for modern LLMs to uncover its under-explored aspects. To achieve this, we dissect LLM unlearning into four essential aspects: formulation, methodologies, evaluation metrics, and applications. We show that there are considerable challenges in both foundational research and practical, use case-driven research. These include: (Generality) A desired solution for LLM unlearning should take into account the generality of the unlearning target and dataset choice, accommodate various model setups including both white-box and black-box scenarios, and consider the specifics of the unlearning method. (Authenticity) LLM unlearning should focus on effectively removing both data influence and specific model capabilities, in order to authenticate unlearning across a range of evaluation methods, particularly in adversarial contexts. (Precision) LLM unlearning should precisely define the scope of unlearning, while ensuring the preservation of general language modeling performance outside this unlearning scope.

By examining the current state of the art, we gain insights for the future development of LLM unlearning. For example, localization-informed unlearning shows promise with possible dual advantages of efficiency and efficacy. Effective unlearning requires careful consideration of data-model influences and adversaries. Despite similarities between LLM unlearning and model editing, they differ in their formulation and methodological design. Furthermore, insights gained from the study of LLM unlearning could catalyze technological advancements in other types of foundation models, *e.g.*, large vision-language models.

In addition, the necessity for regulations or policies to govern unlearning practices is crucial for the future, given the potential implications on privacy, security, and fairness. While existing research has primarily concentrated on auditing unlearning processes related to membership inference, addressing this issue presents an immensely complex challenge. It includes a multitude of factors, including but not limited to data handling/attribution, model governance, transparency, accountability, and verification throughout the unlearning lifecycle. Regulations and policies need to address issues such as data retention, consent management, and the right to be forgotten, which are particularly critical in sensitive domains like healthcare and security applications.

Another future effort we encourage is to build an 'LLM Unlearning Algorithm Card' that carefully details the involved parties, data aimed to be unlearned, evaluation reports, and the implementation details of the unlearning practice.

8 Acknowledgements

S. Liu, J. Jia, and Yuguang Yao were supported by the National Science Foundation (NSF) Robust Intelligence (RI) Core Program Award IIS-2207052 and the Cisco Faculty Research Award. P. Hase and M. Bansal were supported by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, DARPA MCS Grant N66001-19-2-4031, and Google PhD fellowship. We also extend our gratitude to the MIT-IBM Watson AI Lab for their support in this project.

References

Achille, A., Kearns, M., Klingenberg, C., and Soatto, S. Ai model disgorgement: Methods and choices. *arXiv preprint arXiv:2304.03545*, 2023.

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. B. If influence functions are the answer, then what is the question? *Advances in Neural Information Processing Systems*, 35:17953–17967, 2022.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Barbulescu, G.-O. and Triantafillou, P. To each (textual sequence) its own: Improving memorized-data unlearning in large language models. *arXiv preprint arXiv:2405.03097*, 2024.

Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Chowdhury, A. R., Christodorescu, M., Datta, A., Feizi, S., et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.

Becker, A. and Liebig, T. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*, 2022.

Belkadi, L. and Jasserand, C. From algorithmic destruction to algorithmic imprint: Generative ai and privacy risks linked to potential traces of personal data in trained models. In *ICML Workshop on Generative AI + Law*, 2023.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bucknall, B. S. and Trager, R. F. Structured access for third-party research on frontier ai models: Investigating researchers'model access requirements. 2023.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous ai audits, 2024a.

Casper, S., Schulze, L., Patel, O., and Hadfield-Menell, D. Defending against unforeseen failure modes with latent adversarial training, 2024b.

Che, T., Zhou, Y., Zhang, Z., Lyu, L., Liu, J., Yan, D., Dou, D., and Huan, J. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, pp. 4241–4268. PMLR, 2023.

Chen, C., Sun, F., Zhang, M., and Ding, B. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*, pp. 2768–2777, 2022a.

Chen, J. and Yang, D. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*, 2023.

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 499–513, 2022b.

Chen, M., Gao, W., Liu, G., Peng, K., and Wang, C. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7766–7775, 2023a.

Chen, R., Yang, J., Xiong, H., Bai, J., Hu, T., Hao, J., FENG, Y., Zhou, J. T., Wu, J., and Liu, Z. Fast model debias with machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Chien, E., Pan, C., and Milenkovic, O. Efficient model updates for approximate unlearning of graph-structured data. In *The Eleventh International Conference on Learning Representations*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cohen, R., Biran, E., Yoran, O., Globerson, A., and Geva, M. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*, 2023.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Dreyer, M., Pahde, F., Anders, C. J., Samek, W., and Lapuschkin, S. From hope to safety: Unlearning biases of deep models via gradient penalization in latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21046–21054, 2024.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Eldan, R. and Russinovich, M. Who's harry potter? approximate unlearning in llms, 2023.

Fan, C., Liu, J., Hero, A., and Liu, S. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. *arXiv preprint arXiv:2403.07362*, 2024a.

Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024b.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Goland, J. A. Algorithmic disgorgement: Destruction of artificial intelligence models as the ftc's newest enforcement tool for bad data. *Richmond Journal of Law and Technology*, 29(2), 2023.

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Grynbaum, M. and Mac, R. The times sues openai and microsoft over a.i. use of copyrighted work. The New York Times, 2023. URL `https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html`. Accessed: 2024-01-16.

Gu, J.-C., Xu, H.-X., Ma, J.-Y., Lu, P., Ling, Z.-H., Chang, K.-W., and Peng, N. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*, 2024.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Gupta, A., Mondal, D., Sheshadri, A., Zhao, W., Li, X., Wiegreffe, S., and Tandon, N. Editing common sense in transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8214–8232, 2023.

Halimi, A., Kadhe, S., Rawat, A., and Baracaldo, N. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

Hase, P., Bansal, M., Kim, B., and Ghandeharioun, A. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Hase, P., Diab, M., Celikyilmaz, A., Li, X., Kozareva, Z., Stoyanov, V., Bansal, M., and Iyer, S. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *EACL*, 2023b.

He, H., Zha, S., and Wang, H. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*, 2019.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Hoofnagle, C. J., van der Sloot, B., and Borgesius, F. Z. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

Huang, Y., Gupta, S., Xia, M., Li, K., and Chen, D. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Ishibashi, Y. and Shimodaira, H. Knowledge sanitization of large language models. *arXiv preprint arXiv:2309.11852*, 2023.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., and Liu, S. Model sparsity can simplify machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jia, J., Zhang, Y., Zhang, Y., Liu, J., Runwal, B., Diffenderfer, J., Kailkhura, B., and Liu, S. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*, 2024.

Kadhe, S. R., Halimi, A., Rawat, A., and Baracaldo, N. Fairsisa: Ensemble post-processing to improve fairness of unlearning in llms. *arXiv preprint arXiv:2312.07420*, 2023.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24, 2023.

Kumar, V. B., Gangadharaiah, R., and Roth, D. Privacy adhering machine un-learning in nlp. *arXiv preprint arXiv:2212.09573*, 2022.

Kumari, N., Singh, M., Sinha, A., Machiraju, H., Krishnamurthy, B., and Balasubramanian, V. N. Harnessing the vulnerability of latent layers in adversarially trained models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2779–2785, 2019.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Kurmanji, M., Triantafillou, P., and Triantafillou, E. Towards unbounded machine unlearning. *arXiv preprint arXiv:2302.09880*, 2023.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Lermen, S., Rogers-Smith, C., and Ladish, J. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Li, D., Rawat, A. S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., and Kumar, S. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*, 2022.

Li, D., Sun, Z., Hu, X., Liu, Z., Chen, Z., Hu, B., Wu, A., and Zhang, M. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*, 2023a.

Li, M., Davies, X., and Nadeau, M. Circuit breaking: Removing model behaviors with targeted ablation. *arXiv preprint arXiv:2309.05973*, 2023b.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024a.

Li, T. C. Algorithmic destruction. *SMU Law Rev.*, 75:479, 2022.

Li, X., Li, S., Song, S., Yang, J., Ma, J., and Yu, J. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*, 2023c.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

Li, Y., Chen, C., Zheng, X., Liu, J., and Wang, J. Making recommender systems forget: Learning and unlearning for erasable recommendation. *Knowledge-Based Systems*, 283:111124, 2024b.

Liu, B., Liu, Q., and Stone, P. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022a.

Liu, C. Y., Wang, Y., Flanigan, J., and Liu, Y. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*, 2024.

Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. Federated unlearning. *arXiv preprint arXiv:2012.13891*, 2020.

Liu, H., Li, Z., Hall, D., Liang, P., and Ma, T. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023a.

Liu, Y., Fan, M., Chen, C., Liu, X., Ma, Z., Wang, L., and Ma, J. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 280–289. IEEE, 2022b.

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., and Liu, Y. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023b.

Liu, Z., Jiang, Y., Shen, J., Peng, M., Lam, K.-Y., and Yuan, X. A survey on federated unlearning: Challenges, methods, and future directions. *arXiv preprint arXiv:2310.20448*, 2023c.

Lu, X., Welleck, S., Hessel, J., Jiang, L., Qin, L., West, P., Ammanabrolu, P., and Choi, Y. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609, 2022.

Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Madaan, A., Tandon, N., Clark, P., and Yang, Y. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms, 2024.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Menik, S. and Ramaswamy, L. Towards modular machine learning solution development: Benefits and trade-offs. *arXiv preprint arXiv:2301.09753*, 2023.

Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., and Finn, C. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022.

Motoki, F., Pinho Neto, V., and Rodrigues, V. More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*, 2023.

Muresanu, A., Thudi, A., Zhang, M. R., and Papernot, N. Unlearnable algorithms for in-context learning. *arXiv preprint arXiv:2402.00751*, 2024.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Oesterling, A., Ma, J., Calmon, F. P., and Lakkaraju, H. Fair machine unlearning: Data removal while mitigating disparities. *arXiv preprint arXiv:2307.14754*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *ICLR*, 2024.

Pawelczyk, M., Neel, S., and Lakkaraju, H. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Rando, J. and Tramèr, F. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.

Robey, A., Latorre, F., Pappas, G. J., Hassani, H., and Cevher, V. Adversarial training should be cast as a non-zero-sum game. *arXiv preprint arXiv:2306.11035*, 2023.

Sachdeva, B., Rathee, H., Sharma, A., Wydmański, W., et al. Machine unlearning for recommendation systems: An insight. *arXiv preprint arXiv:2401.10942*, 2024.

Sattigeri, P., Ghosh, S., Padhi, I., Dognin, P., and Varshney, K. R. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. *Advances in Neural Information Processing Systems*, 35:35894–35906, 2022.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Si, N., Zhang, H., Chang, H., Zhang, W., Qu, D., and Zhang, W. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*, 2023.

Small, Z. Sarah silverman sues openai and meta over copyright infringement. The New York Times, 2023. URL `https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html`. Accessed: 2024-01-16.

Suliman, M., Kadhe, S., Halimi, A., Leith, D., Baracaldo, N., and Rawat, A. Data forging is harder than you think. In *Privacy Regulation and Protection in Machine Learning*, 2024.

Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., and Ganguli, D. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Thaker, P., Maurya, Y., and Smith, V. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*, 2024.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022a.

Thudi, A., Jia, H., Shumailov, I., and Papernot, N. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4007–4022, 2022b.

Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., Ong, I., Elmaaroufi, K., Abbeel, P., Darrell, T., et al. Tensor trust: Interpretable prompt injection attacks from an online game. *arXiv preprint arXiv:2311.01011*, 2023.

Ullah, E., Mai, T., Rao, A., Rossi, R. A., and Arora, R. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

Wang, B., Zi, Y., Sun, Y., Zhao, Y., and Qin, B. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*, 2024a.

Wang, H., Lin, J., Chen, B., Yang, Y., Tang, R., Zhang, W., and Yu, Y. Towards efficient and effective unlearning of large language models for recommendation. *arXiv preprint arXiv:2403.03536*, 2024b.

Wang, J., Guo, S., Xie, X., and Qi, H. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference 2022*, pp. 622–632, 2022.

Wang, L., Chen, T., Yuan, W., Zeng, X., Wong, K.-F., and Yin, H. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023.

Wang, Y., Wu, R., He, Z., Chen, X., and McAuley, J. Large scale knowledge washing. *arXiv preprint arXiv:2405.16720*, 2024c.

Warnecke, A., Pirch, L., Wressnegger, C., and Rieck, K. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Wen, J., Ke, P., Sun, H., Zhang, Z., Li, C., Bai, J., and Huang, M. Unveiling the implicit toxicity in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Wu, K., Shen, J., Ning, Y., Wang, T., and Wang, W. H. Certified edge unlearning for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2606–2617, 2023a.

Wu, X., Li, J., Xu, M., Dong, W., Wu, S., Bian, C., and Xiong, D. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*, 2023b.

Xu, M., Sun, J., Yang, X., Yao, K., and Wang, C. Netflix and forget: Efficient and exact machine unlearning from bi-linear recommendations. *arXiv preprint arXiv:2302.06676*, 2023.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Yao, J., Chien, E., Du, M., Niu, X., Wang, T., Cheng, Z., and Yue, X. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*, 2024.

Yao, Y., Xu, X., and Liu, Y. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

Yong, Z.-X., Menghini, C., and Bach, S. H. Low-resource languages jailbreak gpt-4, 2023.

Yu, C., Jeoung, S., Kasi, A., Yu, P., and Ji, H. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, 2023.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., and Xu, X. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941*, 2023a.

Zhang, E., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023b.

Zhang, J., Chen, S., Liu, J., and He, J. Composing parameter-efficient modules with arithmetic operations. *arXiv preprint arXiv:2306.14870*, 2023c.

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H., and Li, H. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024a.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024b.

Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pp. 26693–26712. PMLR, 2022.

Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and Liu, S. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023d.

Zhao, K., Kurmanji, M., Bărbulescu, G.-O., Triantafillou, E., and Triantafillou, P. What makes unlearning hard and what to do about it. *arXiv preprint arXiv:2406.01257*, 2024.

Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J., and Chang, B. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*, 2023.

Zhong, Z., Wu, Z., Manning, C. D., Potts, C., and Chen, D. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*, 2023.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.