Open Problems in Machine Unlearning for AI Safety

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**[□], **Jose Hernandez-Orallo**[◇,⋆], **Mor Geva**[◦], **Yarin Gal**[§,‡]

[§]University of Oxford, [♮]Tangentic, [†]University of Tübingen,
[‡]UK AISI, [♠]Massachusetts Institute of Technology, [♡]University of Copenhagen,
[♣]Nanyang Technological University, Singapore AISI, [◇]Universitat Politècnica de València,
[⋆]Leverhulme Centre for the Future of Intelligence, [◦]Tel-Aviv University, [□]Mila - Quebec AI Institute

Abstract

As AI systems become more capable, widely deployed, and increasingly autonomous in critical areas such as cybersecurity, biological research, and healthcare, ensuring their safety and alignment with human values is paramount. Machine unlearning — the ability to selectively forget or suppress specific types of knowledge — has shown promise for privacy and data removal tasks, which has been the primary focus of existing research. More recently, its potential application to AI safety has gained attention. In this paper, we identify key limitations that prevent unlearning from serving as a comprehensive solution for AI safety, particularly in managing dual-use knowledge in sensitive domains like cybersecurity and chemical, biological, radiological, and nuclear (CBRN) safety. In these contexts, information can be both beneficial and harmful, and models may combine seemingly harmless information for harmful purposes — unlearning this information could strongly affect beneficial uses. We provide an overview of inherent constraints and open problems, including the broader side effects of unlearning dangerous knowledge, as well as previously unexplored tensions between unlearning and existing safety mechanisms. Finally, we investigate challenges related to evaluation, robustness, and the preservation of safety features during unlearning. By mapping these limitations and open challenges, we aim to guide future research toward realistic applications of unlearning within a broader AI safety framework, acknowledging its limitations and highlighting areas where alternative approaches may be required.

1 Introduction

For much of the history of machine learning, the primary challenge was enabling models to acquire broad knowledge effectively. However, as models have grown increasingly capable, their ability to access and process vast amounts of information – particularly in sensitive domains such as biology, chemistry, and cybersecurity – has heightened concerns about their potential to cause significant harm. Improving the safety, controllability, and alignment of AI systems increasingly requires preventing them from exhibiting harmful behaviors.

[*]Correspondence to `fazl@robots.ox.ac.uk`. The first and last author rows are the main contributors. Author contributions are detailed in §5.

Preprint. Under review.

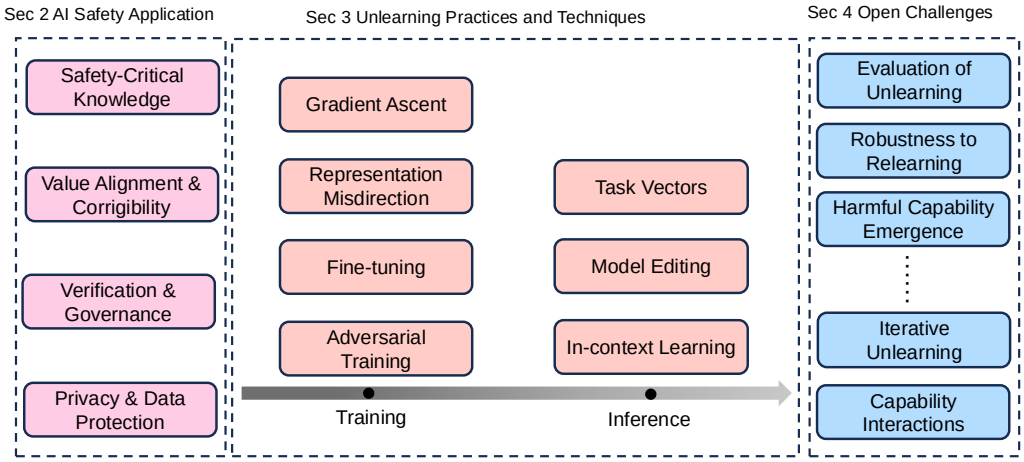

Figure 1: The workflow of our paper. Section 2 discusses applications of unlearning to AI safety, Section 3 surveys methods and practices, and Section 4 outlines open challenges.

An increasingly discussed approach to address these concerns is *machine unlearning*, which is used to remove or suppress unwanted knowledge in AI systems [Cao and Yang, 2015, Bourtoule et al., 2021, Yao et al., 2023, Chen and Yang, 2023, Gu et al., 2024, Maini et al., 2024]. While unlearning shows undeniable promise as an approach to mitigate risks [Jang et al., 2023, Kassem et al., 2023], we argue that fundamental limitations prevent it from being a complete solution to controlling AI capabilities or meeting the objectives necessary for ensuring safety. This has become particularly relevant as AI systems are increasingly used in applications where harmful or inappropriate information poses significant liabilities and demonstrates the possession of safety-critical knowledge [UK AI Safety Institute, 2024].

While existing surveys provide broad overviews of unlearning techniques and applications [Blanco-Justicia et al., 2024, Zhao et al., 2024, Jeong et al., 2024, Liu et al., 2024c], we take a different approach. We critically examine the fundamental constraints that limit unlearning's effectiveness for AI safety applications. We argue that these limitations are particularly severe for capability control when compared to data removal tasks.

We identify four key application areas for unlearning in AI safety. Crucially, we find its effectiveness varies significantly between data removal and capability control tasks across these areas.

- **Safety-Critical Knowledge Management:** Unlearning, when applied to remove harmful knowledge from AI systems, faces a unique challenge - its effectiveness is limited by models' inherent ability to reconstruct capabilities. This limitation is evident even after successful unlearning interventions. For example, even after unlearning of specific chemical synthesis pathways or cybersecurity, models may reconstruct these capabilities by recombining retained benign knowledge, making the unlearning process particularly challenging.
- **Mitigating Jailbreaks:** Unlearning shows promise in removing specific vulnerabilities from training data but faces challenges in preventing broader exploit capabilities. While it can help remove the effects of poisoned data [Goel et al., 2024, Li et al., 2024b] and mitigate risks related to adversarial attacks [Casper et al., 2024b], the fundamental challenge remains that safety bypasses can emerge from necessary system functionalities that cannot be removed.
- **Correcting Value Alignment and Improving Corrigibility:** Current approaches attempt to use unlearning to modify behaviors misaligned with human preference [Ouyang et al., 2022, Wang et al., 2023b] or remove unfair representations. However, we argue that value alignment, being an emergent property of the model's broader knowledge and capabilities, cannot be reliably modified through targeted knowledge removal alone.
- **Privacy and Legal Compliance:** This represents unlearning's most viable application, as it involves removing specific, identifiable data points rather than controlling broader capabilities.

Here, unlearning shows promise as an approach to approximate the removal of training data to comply with regulations like GDPR and CCPA [European Parliament and Council of the European Union, 2016, California State Legislature, 2018, Mantelero, 2013, Voigt and Von dem Bussche, 2017, Hoofnagle et al., 2019].

Challenges and Risks of Unlearning: Achieving effective unlearning relies on accurate identification of the target knowledge [Jia et al., 2024a, Meng et al., 2022], unrecoverable removal of this knowledge [Xhonneux et al., 2024], and comprehensive evaluation on the effect of removal [Maini et al., 2024, Jin et al., 2024] – all of which remain significant technical challenges. Unlearning raises important questions about unintended consequences in contexts like dual-use technologies. The selective removal of knowledge can lead to unintended behaviors within AI models [Liu et al., 2022b]. For example, attempting to remove knowledge of a specific chemical synthesis could unintentionally lead the model to synthesize harmful substances from other materials, using alternative pathways. More generally, the entanglement of knowledge within AI models makes it difficult to predict the potential downstream effects of unlearning, especially its impact on other knowledge, necessitating thorough evaluation and testing to minimize risks. These limitations include unexpected interactions between different safety measures, such as how unlearning can interfere with model robustness and how combining multiple safety techniques can lead to compounding performance degradation.

Contributions: This paper critically examines the limitations of machine unlearning for AI safety applications, with a particular focus on the fundamental constraints that prevent it from being a complete solution for capability control. We provide an overview of current unlearning techniques and their evaluation methods. We then identify and analyze key open problems that limit unlearning's effectiveness for AI safety, including the emergence of harmful capabilities from benign knowledge, challenges in dual-use contexts, and difficulties in verification. By mapping these fundamental limitations, we aim to guide future research toward realistic applications of unlearning within a broader safety framework, while highlighting areas where alternative approaches are needed.

2 Applications of Unlearning for AI Safety: An Overview

To fully explore the interaction between unlearning and AI safety, it is essential to first clarify the objective of machine unlearning and how machine unlearning modifies system behavior for the broader landscape of AI safety.

Goal: Machine Unlearning for AI Safety

Machine unlearning aims to modify an AI system so it *forgets* specific knowledge or behaviors, examples of which are provided in a "unlearning corpus". *Forgetting* means the updated system should no longer exhibit or retain any knowledge or behaviors demonstrated in the forget corpus. Simultaneously, the system's performance on tasks unrelated to the forget corpus must remain unaffected, ensuring its overall utility is preserved.

While privacy-focused indistinguishability objectives seek to prevent models from learning private information with minimal impact on capabilities [Yu et al., 2021, Charles et al., 2024], unlearning aims to *undo* the effect of learning from certain data points entirely. In other words, unlearning for AI safety requires the deliberate suppression or removal of specific knowledge or behaviors, ensuring the system cannot express or leverage the forgotten information. In this section, we shall provide a more in-depth exploration of the application areas of unlearning for AI safety: (i) Safety-Critical Knowledge Management, (ii) Correcting Value Alignment and Improving Corrigibility, (iii) Verification and Governance Aspects, and (iv) Privacy and Legal Compliance.

2.1 Unlearning for Safety-Critical Knowledge Management

Dual-Use Hazardous Knowledge: Increasingly advanced models may pose serious real-world risks from hazardous knowledge such as assisting in synthesis pathways for CBRN materials or cybersecurity vulnerabilities [Hendrycks et al., 2023]. Current approaches address this through multiple strategies: knowledge editing and targeted pruning to identify and remove specific parameters encoding undesired knowledge [Meng et al., 2022, 2023]; analysis of model attention weights and activation patterns to identify neurons strongly associated with targeted knowledge [Wu et al., 2023,

Farrell et al., 2024, Dai et al., 2022]; and comprehensive testing on held-out datasets to verify the removal of hazardous knowledge [Deeb and Roger, 2024, Li et al., 2024a].

Mitigating Adversarial Attacks & Jailbreaks: Models face vulnerabilities from carefully crafted inputs that can bypass safety measures. For instance, adversarial prompts might cause language models to output inappropriate violent, sexual, or biased content [Wei et al., 2023]. Protection against these threats could involve selective forgetting of features associated with adversarial examples, training with weighted adversarial examples to reduce attack pattern sensitivity [Casper et al., 2024b, Sheshadri et al., 2024], and rigorous verification through adversarial robustness testing [Isonuma and Titov, 2024].

2.2 Correcting Value Alignment and Improving Corrigibility via Unlearning

Reward Hacking: In reinforcement learning contexts, policy models can discover unintended shortcuts or loopholes in reward models [Skalse et al., 2022]. A common example is how language models may generate unnecessarily verbose outputs after preference learning, exploiting the tendency for longer responses to be preferred in pair-wise preference datasets [Park et al., 2024, Shen et al., 2023, Singhal et al., 2023]. Popular approaches that aim to address this include: early stopping, information-theoretic reward modeling [Miao et al., 2024], spurious factor disentangling [Chen et al., 2024], and implementing constrained optimization with heuristic human knowledge about the shortcut [Meng et al., 2024].

Value Alignment: Models may develop behaviors that are misaligned with specific human preferences [Ouyang et al., 2022]. Potential directions to achieve better value alignment include: reward modeling with iterative refinement based on human feedback, applying gradient-based modification of parameters associated with misaligned behaviors [Jang et al., 2023, Gu et al., 2024], and integrating interpretability methods for verification [Vidal et al., 2024, Hong et al., 2024].

Situational Awareness: Recent LLMs have demonstrated some awareness of the situations of their own development, including knowledge about their creators, capabilities, and the broader LLM development and deployment lifecycle [Laine et al., 2024, Perez et al., 2023]. With greater situational awareness, some researchers have raised concerns that AI systems could leverage this knowledge harmfully if their goals are misaligned with the goals of their developers [Ngo et al., 2024, Carlsmith, 2023, Krasheninnikov et al., 2024]. For example, AI systems could manipulate the known biases of human reviewers to earn higher reward scores [Sharma et al., 2023, Williams et al., 2024, Wen et al., 2024a], or insert security vulnerabilities into users' codebases to enable self-exfiltration [Meinke et al., 2024]. Recent work shows that in certain circumstances, large language models can leverage situational knowledge about their training process to "fake alignment" in an effort to prevent a developer from changing the model's objectives [Greenblatt et al., 2024]. New unlearning methods could help address these risks by providing a mechanism to selectively modify or remove an AI system's knowledge of its own situation. This controlled modification serves as both a diagnostic tool, revealing potential misbehavior patterns, and a preventive measure, allowing misbehavior to be corrected before deployment.

2.3 Verification and Governance Aspects of Unlearning

As AI systems become increasingly pervasive, ensuring compliance with governance frameworks becomes critical [de Almeida et al., 2021]. Strategies for verification and compliance include: explicit compliance objectives in system design, utilizing interpretability methods to identify non-compliant behaviors [Casper et al., 2024a], integrating formal verification methods where applicable [Xu et al., 2024a], and ensuring adherence to industry standards and regulatory requirements [UKAISI, USAISI]. Designing and implementing robust evaluation standards [Maini et al., 2024, Jin et al., 2024], algorithms for formal verification [Xu et al., 2024a, Zhang et al., 2024], and audit protocols with varying levels of model access are essential components of effective machine unlearning. For example, Hong et al. [2024] identity parametric concept vectors that are strongly correlated with specific dual-use knowledge and assesses the effectiveness of the unlearning approach by the alternation of concept vectors.

2.4 Unlearning for Privacy, Data Protection and Legal Compliance

In today's digital landscape, apps, websites, and home automation devices continuously collect user behavior patterns, including personally identifiable information (PII). When users become aware of potential PII leakage, they may request deletion as mandated by regulations like GDPR [European Parliament and Council of the European Union, 2016]. However, simply deleting PII from databases proves insufficient, as models may have memorized this information during training, potentially allowing for data reconstruction by model developers or even other users of a model. Existing verification techniques can only provide probabilistic guarantees about knowledge removal. This mismatch creates practical challenges where legal frameworks assume binary deletion while technical reality operates on a continuous spectrum.

To fully address these privacy concerns, several key capabilities are necessary. First, unlearning algorithms must be able to identify and remove specific factual knowledge about individuals. Second, they need mechanisms to modify model behavior to avoid generating content that reveals private information. Finally, they can adjust internal representations to reduce membership inference risks [Sula et al., 2024]. While this application represents an essential aspect of AI safety, it will not be the focus of our study, as it has been the primary focus of existing unlearning surveys (see Yao et al. [2024b]),

3 Unlearning Practices and Evaluation Techniques for AI safety

In pursuit of a safe and reliable AI system, various techniques have been developed for machine unlearning. In this section, we provide an overview of machine unlearning including current techniques for unlearning (Section 3.2) and evaluation dimensions to assess unlearning algorithms (Section 3.1).

3.1 Evaluation Methods for Unlearning: What are the promising directions for AI Safety?

Establishing evaluation methods and metrics is crucial to assess and compare the effectiveness of unlearning techniques. Drawing inspiration from recent literature on machine unlearning and interpretability, we can identify several key evaluation metrics and methods.

Generalizability: Downstream performance is an important dimension to evaluate the success of unlearning. If the targeted objective of unlearning is a specific down-stream task [Wang et al., 2023a, Pawelczyk et al., 2024, Ishibashi and Shimodaira, 2023], the down-stream performance on the specific task is the most direct way for evaluation. For example, ZRF score [Chundawat et al., 2023a] measures the similarity between the unlearned model performance and a randomly initialized model. On the other hand, if the unlearning objective is a specific type of harmful knowledge, evaluation on corresponding harmfulness benchmarks is indispendisble [Gehman et al., 2020, Lin et al., 2023, Parrish et al., 2022, Li et al., 2024a].

Locality: Locality is another important dimension for unlearning algorithms. During the unlearning process, the benign capacities of the AI system is supposed to remain unaffected. For unlearning on large language models, language modeling benchmarks [Merity et al., 2016] are good choices to measure the maintenance of basic language modeling ability. Meanwhile, instruction-following benchmarks [Dubois et al., 2023, Zheng et al., 2023, Zhou et al., 2023] and world knowledge benchmarks [Hendrycks et al., 2021, Wang et al., 2024c] are widely adopted to measure the side effects of unlearning on the instruction-following and commonsense reasoning abilities of LLMs.

Efficiency and Resource: Aside from downstream accuracy and behavioral change, practical considerations are essential for real-world applications of unlearning. Time cost and computational resources required are two important factors when comparing and assessing different unlearning algorithms, since some approaches to unlearning involve the computation or approximation of the Hessian matrix [Jia et al., 2024b, Gu et al., 2024], which is time-consuming. In addition, early work on the unlearning of the vision model tends to track the time or the difficulty involved in relearning the unlearned task [Tarun et al., 2024, Chundawat et al., 2023b, Lo et al., 2024] which provides insights into the depth of unlearning.

Table 1: A summary of evaluation dimensions for unlearning.

Assess Dimension	Requirement
Generalizability	Whether the effect of unlearning is generalizable to other forms of expression for the forgetting corpus.
Locality	Whether the benign capacities of the AI system remain unaffected.
Efficiency and Resource	Whether the unlearning algorithm could be used with limited computation and time.
Robustness to Adversarial Attack	Whether the unlearning is robust to membership inference and inversion attacks.
Robustness to Relearning	Whether the unlearned knowledge is difficult to be restored and relearned.

Robustness to Adversarial Attack: For traditional applications and scenarios of unlearning where user privacy and data security is a primary concern, the model vulnerability when faced with an adversarial attack can serve as an optional metric to evaluate the effectiveness of unlearning and whether it contributes to the robustness of the language models. To achieve the goal, there are two types of particular attacks of interest, namely membership inference attack [Mattern et al., 2023, Duan et al., 2024a] and model inversion attack [Nguyen et al., 2023, Morris et al., 2024]. Membership inference attack aims to determine whether a given datum is included in the training set, mostly relying on the model likelihood [Shi et al., 2024a, Mattern et al., 2023] on the datum. Model inversion attack, on the other hand, targets reconstructing the training data from the model. In spite of their potential to provide more in-depth understanding of the relationship between unlearning and robustness, how to implement effective model membership attack and model inversion attack on large language models remain an unsolved problem [Duan et al., 2024b].

3.1.1 What Metrics Would be Most Suitable for AI Safety Applications?

These simple metrics related to unlearning success are useful, standardizable measures for unlearning progress. However, they are not suitable for real-world safety circumstances in which unlearning may be relied on. We expand here on on what we feel are the most promising directions for metrics.

There are many ways to elicit knowledge from models [Shi et al., 2024b, Lynch et al., 2024, Hayes et al., 2024, Liu et al., 2024c], so placing too great a focus on some measures can cause other practical failure modes to be neglected. This highlights the need for adversarial evaluations for machine unlearning to more thoroughly evaluate the practical success of unlearning methods [Goel et al., 2022, Liu et al., 2024a]. Prior work has shown that input-space knowledge elicitation techniques can be effective at extracting unlearned knowledge from LLMs [Lynch et al., 2024, Shumailov et al., 2024, Łucki et al., 2024]. Other threat models may also be applicable for cases in which models might be deployed open-source or with a fine-tuning API. Patil et al. [2023], Lynch et al. [2024], and Hong et al. [2024] have all demonstrated that "unlearned" knowledge can be extracted from analysis of the internal mechanisms of LLMs.

However, a variety of works have shown that unlearned knowledge can be very sample-efficiently re-learned through few-shot fine-tuning [Hu et al., 2024, Yuan et al., 2024, Henderson et al., 2023, Tamirisa et al., 2024, Sheshadri et al., 2024, Lo et al., 2024]. This poses interesting open challenges to suitable evaluation metrics for unlearning in safety-specific applications.

3.2 Current Unlearning Methods

To achieve effective unlearning for AI systems, various approaches have been developed in recent years [Jang et al., 2023, Ilharco et al., 2023, Ishibashi and Shimodaira, 2023, Pawelczyk et al., 2024]. In this subsection, we provide an extensive review of recent approaches and elaborate on their pros and cons.

Gradient Ascent: To unlearn or offset the effect of the unlearning corpus, an intuitive approach is to reverse the direction of the parameter gradient and maximize the loss function on the unlearning

Table 2: A summary of current approaches for unlearning.

Method	Advantage	Limitation	Example
Gradient Ascent	straightforward and easy to implement	unlearning failure and catastrophic collapse	SOUL [Jia et al., 2024b]
Task Vector	straightforward and easy to implement	sensitive to hyper-parameter	Task Arithmetic [Ilharco et al., 2023]
Model Editing	light-weighted and minor intervention	hard to locate relevant representations	DEPN [Wu et al., 2023]
Representation Misdirection	light-weighted and retains performance	non-robust to adversarial finetuning	RMU [Huu-Tien et al., 2024]
Adversarial training	robust to model modification	less efficient	LAT [Sheshadri et al., 2024]
Finetuning on curated data	better controllablity	require constructing new data	Knowledge Sanitization [Ishibashi and Shimodaira, 2023]
In-context learning	black-boxed and API-friendly	high computation cost for in-context demonstrations	In-context Unlearning [Pawelczyk et al., 2024]

corpus [Jang et al., 2023]. Mathematically, the learning objective of gradient ascent to maximize is

$$\mathcal{L} = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_f} \log \mathcal{M}(\boldsymbol{y} \mid \boldsymbol{x}) \tag{1}$$

However, first-order gradient ascent on the full set of model parameters tends to suffer from performance degradation, as the reversed gradient disrupts not only knowledge and abilities related to the unlearning corpus but also the irrelevant knowledge [Wang et al., 2024b, Zhao et al., 2024]. To deal with the deficits of first-order gradient ascent, a common approach is to balance the gradient ascent with KL divergence [Wang et al., 2023a, Yao et al., 2024a, Chen and Yang, 2023] or language modeling loss on the remaining corpus [Maini et al., 2024, Yu et al., 2023]. For instance, Wang et al. [2023a], Chen and Yang [2023] minimizes the KL divergence between the unlearned model $\mathcal{M}'$ and the original model $\mathcal{M}$ on the remaining corpus while enlarging their KL divergence on the unlearning corpus. Apart from constraining the model parameter updates with the KL divergence, performing gradient ascent only on the model parameters that are most related to the unlearning target [Wang et al., 2024b, Wu et al., 2023, Yu et al., 2023] is another plausible approach. As an example, Yu et al. [2023] identify bias-related neurons with integrated gradient [Sundararajan et al., 2017] and conduct gradient ascent only on the selected neurons. Recently, second-order optimization with Sophia [Liu et al., 2023] or inverse empirical Fisher approximation seems to be another promising approach to implement gradient ascent without sacrificing the model utility [Jia et al., 2024b, Gu et al., 2024].

Representation Misdirection: Different from gradient ascent that optimizes model parameters towards a reversed optimization function, representation misdirection aims to remove the unlearning target by misdirecting their intermediate hidden representation towards random noise. For example, Li et al. [2024a] minimize the Euclidian distance between the representation of CBRN knowledge after the eighth transformer layer and a fixed random noise. Similarly, Rosati et al. [2024b] and Zou et al. [2024] optimize the representation of potentially harmful knowledge towards an uninformative noise. However, a recent work [Huu-Tien et al., 2024] suggests that the techniques might fail when the norm of the harmful representation is larger than that of the random noise. They further propose to use an adaptive noise for scaling the random noise to avoid this case. Meanwhile, another recent work [Andy] finds that a significant proportion of the effect of representation misdirection is attributable to the random noise injected into the residual norm when the harmful context is detected, rather than removing the potentially harmful knowledge directly.

Task Vectors: Recently, task vectors have emerged as a lightweight technique to achieve unlearning. Originally proposed by Ilharco et al. [2023], task vectors refer to the difference between a fine-tuned model and its base pre-trained model in the model parameter space. To be more specific, to achieve machine unlearning, first a reinforced model $\mathcal{M}_f$ is obtained by tuning the original language model $\mathcal{M}$ on the unlearning corpus. Then we can find the forgetting task vector by subtracting $\mathcal{M}$ from $\mathcal{M}_f$ in an element-wise manner. Subsequently, the task vector is subtracted from the original model to produce an unlearned model, which is verified to forget the learning corpus without harming other abilities [Ilharco et al., 2023, Zhang et al., 2023a]. Intuitively, the task vector serves as the

aggregation of gradient effects caused by the unlearning corpus $\mathcal{D}$ and it can therefore be "deleted" in parameter space. Following Ilharco et al. [2023], Zhang et al. [2023a] extends the paradigm to parameter-efficient fine-tuning of large language models and verifies its efficacy on toxicity reduction. Similarly, Barbulescu and Triantafillou [2024], Ni et al. [2024], Liu et al. [2024b] train an unlearning task vector on the target downstream task, toxic corpus, or out-of-date knowledge to obtain an unlearning task vector and subtract it from the original model.

Model Editing: Different from gradient ascent or model merging which directly alter the model parameter, model editing modifies the intermediate hidden state or the logits to change the model behavior, usually with the help of interpretability tools [Huben et al., 2024, Koh and Liang, 2017]. For example, Wu et al. [2023] identify the neurons in the model that contribute most to the model prediction of privacy-related content through integrated gradient [Sundararajan et al., 2017] and the activations from the neurons are masked to prevent the generation of user privacy. Farrell et al. [2024] use sparse auto-encoder to manipulate the activation of monosemantic features related to the knowledge to forget. In a similar vein, Guo et al. [2024] localize the fact lookup stage [Nanda et al., 2023] of language models to specific MLP layers and keep other parameters unchanged. Apart from word embedding and intermediate neuron action, the editing can be applied to output logits as well. For example, Huang et al. [2024a] compare the logit difference of a pair of expert and amateur models and directly applies the offset in logits to a large black-boxed model. To some extent, contrastive decoding can be included as a kind of model editing and there is a surge of studies that apply contrastive decoding to unlearn the hallucinated and unsafe content [Zhang et al., 2023b, Zhong et al., 2024].

Finetuning on Curated Data: Another approach to eliminating the effect of unlearning corpus $\mathcal{D}_f$ is to substitute the parametric knowledge to be forgotten with an insensitive one, implemented by finetuning the model on some curated data. Typically, the new data for fine-tuning is modified unlearning corpus $\mathcal{D}_f$ and there are two commonly used strategies to modify the unlearning corpus. One strategy is curating refusal data [Ishibashi and Shimodaira, 2023]. Specifically, refusal data pairs sensitive user queries with a fixed refusal sentence like "I don't know". Consequently, model fine-tuned on the refusal data learns to refuse to answer questions about the unlearning corpus. Another strategy is to construct anonymized data, by replacing the key entities and terms in the unlearning corpus with pre-defined alternatives. For example, Eldan and Russinovich [2023] construct a dictionary mapping every named entity in Harry Potter to an anonymized counterpart and [Yao et al., 2024a] generate anonymized data based on unlearning corpus by masking out the sensitive tokens.

Adversarial Training Against Weight-Space and Latent-Space Attacks: A challenge of fine-tuning-based methods is the limited ability of fine-tuning to 'deeply' remove capabilities from a model in a way that is resistant to resurfacing from anomalies, attacks, or model modifications [Shayegani et al., 2023, Qi et al., 2023, Bhardwaj and Poria, 2023, Qi et al., 2024, Hu et al., 2024, Ji et al., 2024, Wei et al., 2024]. This challenge has motivated work to more rigorously remove undesirable knowledge from models by training them to unlearn in a way that is robust to model manipulations. This work has included adversarial training on input- [Tarun et al., 2023] or latent-space [Casper et al., 2024b] attacks which have been used to more effectively remove unlearned knowledge [Zeng et al., 2024, Yuan et al., 2024, Huang et al., Sheshadri et al., 2024]. Others have trained models under adversarial weight-space perturbations to the same effect [Henderson et al., 2023, Deng et al., 2024, Tamirisa et al., 2024, Huang et al., 2024b]. However, research on these methods is still nascent, and they are relatively inefficient compared to fine-tuning alone.

In-Context Learning: Since the cost for fine-tuning or post-training may go beyond the computation budget for a large proportion of academic institutes, prior studies investigate in-context learning [Pawelczyk et al., 2024, Thaker et al., 2024, Muresanu et al., 2024] as a computationally friendly substitution for fine-tuning or post-training. Specifically, by instructing language model to avoid using particular knowledge when answering user queries [Thaker et al., 2024] or inserting demonstrations from forgetting corpus with their labels flipped [Pawelczyk et al., 2024], in-context learning provides a light-weight solution for unlearning. But counter-intuitively, the computation cost of in-context learning can be even larger than fine-tuning [Liu et al., 2022a] since we have to recompute the representation of the in-context demonstrations if we reselect in-context demonstrations for every new query from users. In addition to complementing system prompts with unlearning

instructions or demonstrations at inference time, in-context learning plays an important role in constructing moderation API and content classifier. Prompted with legal rules or social norms [Xu et al., 2023], the API or classifier can filter out unsafe queries or generations [Inan et al., 2023, Hurst et al., 2024], though the parametric knowledge of the language model remains unchanged.

4 Open Problems in Machine Unlearning for AI Safety

The application of machine unlearning to AI safety introduces unique challenges that extends beyond traditional machine unlearning. Resolving these open problems will be crucial for developing unlearning as a reliable tool for AI safety.

4.1 Evaluation of Unlearning

Simple metrics that check whether models can reproduce specific training examples fail to capture the deeper challenges of unlearning in safety-critical contexts. When models undergo modifications, face adversarial attacks, or encounter unusual inputs, unlearned capabilities can unexpectedly resurface - particularly in cases where the unlearning relied on fine-tuning or basic parameter adjustments [Hu et al., 2024, Łucki et al., 2024, Deeb and Roger, 2024]. This happens because these methods typically mask rather than eliminate capabilities, leaving the fundamental neural patterns that enable them largely untouched [Jain et al., 2023]. More rigorous standards are helpful in addressing these limitations. This includes ensuring that forgotten knowledge cannot be recovered, does not reappear during extended interactions, and remains inaccessible even in new contexts or under adversarial pressure. Such verification must consider both specific knowledge removal and broader capability prevention [Jin et al., 2024, Maini et al., 2024], examining:

1. How models might rebuild unlearned capabilities through indirect means, like reconstructing security exploits by combining basic programming concepts
2. Ways that remaining knowledge could combine to recreate harmful capabilities
3. Whether the unlearning remains robust against deliberate attempts to recover removed capabilities

These challenges highlight a fundamental issue: preventing a model from reproducing specific content doesn't guarantee that it can't reconstruct the underlying capabilities through other means. Developing reliable verification methods that can detect and prevent such reconstruction remains an active challenge in the field.

Evaluation Challenges: As discussed earlier in Section 3.1, existing metrics and assessments appear successful on specific tasks, but they might fail to capture the broader impact on model capabilities - a model might pass targeted tests while retaining problematic behaviors that appear in subtler ways. Long-term effects remain particularly challenging to assess. Current metrics typically focus on immediate post-unlearning evaluation but provide little insight into how unlearning impacts model behavior over extended periods or under adversarial attack [Chen et al., 2021]. This limitation becomes critical when considering how models might gradually reconstruct supposedly removed capabilities through continued operation. The field needs standardized benchmarks that the AI safety community can rely on. While newer approaches show promise—such as tracking changes in model activations and using interpretability tools [Belrose et al., 2023, Bricken et al., 2023, Foote et al., 2023] to understand how unlearning affects learned features—integrating these into practical, computationally feasible frameworks remains an open challenge.

4.2 Robustness to Relearning

Even when effective, unlearning can be surprisingly vulnerable to fine-tuning and could quickly relearn the hazardous knowledge [Lo et al., 2024, Lynch et al., 2024, Deeb and Roger, 2024], even if fine-tuned on small amount of benign, unrelated data [Łucki et al., 2024, Hu et al., 2024]. This suggests that existing techniques have a limited ability to thoroughly remove hazarous knowledge from LLMs. It also poses a significant challenge to the safety of open-source models or proprietary models that can be fine-tuned [Achiam et al., 2023]. Some works have aimed to perform unlearning in a way that is more robust to post-deployment tampering [Deng et al., 2024, Henderson et al., 2023,

Huang et al., 2024c, Rosati et al., 2024b,a, Tamirisa et al., 2024]. However, these existing methods suffer from major tradeoffs with efficiency, stability, and performance on benign tasks. Establishing benchmarks and improving techniques for tamper-resistant unlearning is an ongoing challenge.

4.3 Dual-use Capabilities Can Emerge From Beneficial Elements

The distinction between knowledge and capabilities presents a fundamental challenge in safety-critical knowledge removal domains [Li et al., 2024a]. Traditional unlearning scenarios target specific data points or patterns, but safety-critical applications must focus on preventing dual-use capabilities while preserving beneficial ones [Jin et al., 2024]. While knowledge typically represents localized information (like specific facts or patterns), capabilities emerge from the complex interaction and integration of multiple pieces of knowledge potentially across different domains, making them inherently more distributed throughout a model's parameters and allowing potentially harmful capabilities to emerge even from combinations of seemingly harmless knowledge. It remains a critical challenge to prevent harmful capabilities from emerging from combinations of knowledge.

4.4 Context-Dependent Challenges

Knowledge removal in AI systems presents distinct challenges depending on how the knowledge is represented and used. At the simplest level, removing the effect of specific training data points is relatively straightforward. More complex cases arise when removing general knowledge that may be distributed across multiple training instances. The most challenging scenarios involve knowledge that is either context-dependent (as in dual-use cases) or knowledge that combines with other model capabilities to enable complex behaviors not directly tied to specific training examples – for example, when basic language understanding combines with reasoning skills to enable sophisticated problem-solving abilities.

This progression reflects an increasing difficulty in attribution - from clearly identifiable training data points, to distributed but traceable knowledge, to knowledge whose safety depends on the context, and finally to capability-related knowledge where attribution becomes highly complex due to how capabilities arise from the synthesis of multiple basic competencies rather than from discrete, identifiable training examples.

The management of dual-use knowledge in domains like CBRN and cybersecurity necessitates sophisticated access control mechanisms. Knowledge appropriate for trusted contexts may prove dangerous in others—for instance, detailed vulnerability information crucial for cybersecurity professionals requires careful containment. The challenge extends beyond simple knowledge removal to implementing context-dependent access controls through unlearning techniques [Anonymous, 2024, Cui et al., 2024].

In addition, removing dual-use knowledge can create blind spots in the model's safety mechanisms. For instance, removing a detailed understanding of security vulnerabilities might prevent a model from effectively identifying and avoiding similar vulnerabilities in new contexts. This creates a practical dilemma: maintaining robust safety guardrails may require preserving some of the very knowledge we aim to remove [Longpre et al., 2023].

4.5 Neural and Representational Level Interventions

Attempts to manipulate knowledge in AI systems must grapple with the complex relationship between low-level neural parameters (weights and activations) and higher-level patterns that emerge from them. While we can intervene at the level of individual weights or try to target specific semantic concepts [Wu et al., 2023, Yu et al., 2023, Meng et al., 2022], the relationship between these different aspects of the system is not fully understood. For example, removing a model's ability to generate harmful content might inadvertently affect its understanding of context and nuance in related but benign domains.

These challenges directly connect to our earlier discussion of knowledge and capabilities in Section 4.4. While knowledge might be modified at the representational level (like removing understanding of specific harmful concepts), the distributed nature of capabilities means that the corresponding neural-level changes could affect multiple capabilities simultaneously.

The interaction between these levels becomes particularly significant in the context of the context-dependent challenges discussed earlier in Section 4.4. This coordination between neural-level modifications and representational-level changes must ensure that protective mechanisms remain robust across different contexts while preserving the model's core functionalities.

This coordination requires advances in three key areas, namely casual interventions, predictive safety frameworks and validation, and monitoring systems. Specifically, causal intervention techniques is the basis for (i) modifying specific computational circuits while detecting and limiting effects on connected circuits; (ii) creating robust methods to handle distributed features in superposition while preserving their independent function; (iii) building tools that can precisely target modifications at different scales (neuron-level to network-level) while maintaining circuit integrity. Meanwhile, the predictive safety framework encompasses (i) constructing predictive models that can estimate how modifications will propagate through the network's computational graph; (ii) developing formal verification methods to ensure modifications stay within safety bounds; (iii) creating techniques to identify potential interactions between modified circuits and seemingly unrelated capabilities. Validation and monitoring systems require (i) building comprehensive testing frameworks that can detect subtle changes in both targeted and untargeted capabilities; (ii) implementing continuous monitoring systems that can track the long-term effects of modifications; and (iii) developing methods to validate whether modified circuits maintain their intended function under different contexts and inputs.

Each of these areas needs solutions that work both at the level of individual neurons and across the broader network to ensure reliable and controlled modifications.

4.6 Continual and Iterative Unlearning

Current unlearning methods mostly have a fixed unlearning corpus and struggle to fully unlearn unwanted behaviors and preserve unrelated knowledge at the same time. As a possible remedy, iterative unlearning has the potential for attaining a better trade-off among multiple unlearning objectives. By dynamically adjusting the weights of multiple optimization objectives in the loss function of unlearning [Yu et al., 2023], iterative learning enables a balance between effectiveness and locality.

Sequential unlearning is another common challenge in real-world scenarios where users make sequential unlearning requests across time [Wang et al., 2024a, Hartvigsen et al., 2023, Mitchell et al., 2022]. Current unlearning methods show trade-offs between the degree of unlearning achieved and preservation of model utility. These methods typically achieve partial unlearning while experiencing some degradation in model performance. When unlearning requests are processed sequentially, each operation begins with a model that is already degraded, leading to a cumulative loss in utility over time. As performance deteriorates over multiple rounds, the model can be destabilized and most valuable knowledge can be unintentionally erased. One key direction is developing solutions that can scale across large number of unlearning requests, maintaining the broad knowledge of foundation models measured by performance, while accommodating numerous unlearning requests across time.

The specification of the unlearning targets themselves requires an iterative approach. Similar to how reward specification in reinforcement learning often requires iterative refinement when agents find unexpected ways to exploit the initial reward function [Amodei et al., 2016, Krakovna et al., 2020, Gajcin et al., 2023], precisely defining what knowledge to unlearn often requires multiple rounds of adjustment as we discover new ways that harmful capabilities can manifest. This necessitates developing frameworks for progressive refinement of unlearning targets, allowing more nuanced and precise interventions over time, aimed at reducing the risk of missing critical information or removing valuable data. Additional key questions include whether unlearning operations should be amortized by combining multiple unlearning requests before execution, and how to effectively intersperse unlearning with ongoing learning processes.

4.7 Capability Interactions and Dependencies

The interactions between safety measures reveal several challenges and opportunities that emerge when implementing unlearning in practice. These relationships go beyond simple dependencies, exposing fundamental tensions and trade-offs to be carefully managed.

The relationship between robustness and knowledge management reveals complex challenges. Attempts to increase model robustness through selective forgetting can expose unexpected knowledge dependencies. For example, when unlearning CBRN knowledge, models may develop new synthesis pathways that are harder to detect, suggesting that knowledge representations are more deeply entangled than previously understood. Removing specific capabilities could affect seemingly unrelated tasks, while models may develop compensatory behaviors that create new safety risks [Lo et al., 2024]. This observation indicates that the granularity of behaviour removal needs to match the granularity of knowledge representation.

The push for automated safety mechanisms has revealed fundamental limits in unlearning capabilities. Fully automated unlearning systems consistently struggle to balance precision of knowledge removal against computational efficiency, maintenance of model utility against safety guarantees, and generalization of safety rules against context-specific requirements [Wang et al., 2024a]. These trade-offs suggest that hybrid approaches combining automated detection with human oversight may be necessary, despite their inherent scalability challenges.

Perhaps most intriguingly, the interaction between different safety measures can produce unexpected emergent behaviors. Models that undergo repeated cycles of unlearning and retraining may develop resistance to future modification attempts. Safety mechanisms designed for one domain may create vulnerabilities in others, and the combination of multiple safety techniques can lead to compounding performance degradation. These emergent properties suggest that the interactions between safety measures are more complex than previously recognized.

These observations point to several research directions. These include developing methods for mapping and managing knowledge dependencies before unlearning. New verification frameworks could be created to align technical capabilities with regulatory requirements, while hybrid unlearning architectures needs to be designed that effectively balance automation with human oversight. Additionally, techniques for predicting and managing emergent behaviors in safety systems are becoming increasingly appealing for risk management.

Beyond unlearning, there are many other techniques to modify model behavior, and they vary in their implementation requirements. Some can be automated with minimal human intervention, such as privacy protection mechanisms that automatically detect and remove memorized data [Wen et al., 2024b], basic adversarial defense systems operating through pattern recognition [Feng et al., 2023, Liu et al., 2022c], and standardized compliance checks [Xu et al., 2024b]. Others might require significant human oversight, particularly in areas such as value alignment judgments, identification of harmful knowledge in context, and complex regulatory compliance decisions. Understanding these requirements helps inform how and when unlearning techniques can be most effectively applied.

5 Conclusion

In this paper we have explored the usefulness of unlearning techniques for AI Safety. Despite its popularity as a potential solution for AI Safety challenges, a series of fundamental constraints limit unlearning's effectiveness, particularly for capability control. While unlearning shows promise for data removal tasks, our analysis reveals inherent limitations in controlling AI capabilities. The emergent nature of dual-use capabilities from combining seemingly benign knowledge presents an inherent limitation that unlearning cannot fully address. The context-dependent nature of knowledge representation, especially in dual-use scenarios, further prevents reliable selective capability removal. Several constraints compound these core limitations. Current approaches to neural-level interventions often produce unintended effects on broader model capabilities, adding practical challenges to selective capability control, while the difficulty of verifying unlearning success and robustness against relearning raises additional concerns. Furthermore, unlearning interventions can create tensions with existing safety mechanisms, potentially affecting their reliability. These inherent constraints demonstrate that unlearning must be viewed as one component in a broader safety framework, not a complete solution. By mapping these fundamental limitations, we aim to guide research toward developing realistic approaches that acknowledge unlearning's bounded role in AI safety. Future work should proceed along two paths: (1) identifying specific use cases where unlearning can be effectively applied for data removal, and (2) developing alternative approaches for capability control given the fundamental limitations of unlearning in this domain.

Author Contribution

Fazl Barez conceptualized and led the project, wrote the majority of the manuscript, and managed its overall development. **Tingchen Fu** and **Ameya Prabhu** made substantial contributions to sections 2 and 3, and significantly improved the manuscript through extensive editing and revision.

Stephen Casper, **Adel Bibi**, **Aidan O'Gara**, and **Robert Kirk** contributed with technical feedback and helped with copyediting parts of the paper which helped refine the paper's positioning.

David Kruger, **Sören Mindermann**, **Jose Hernandez-Orallo**, **Mor Geva**, and **Yarin Gal** provided detailed feedback and advice throughout the project.

Acknowledgments

We thank Yoshua Bengio, Andy Zou, Katrina Dickson, Niloofar Mireshghallah, Anna Ritchie, Xander Davies and the Tangentic AI team for their comments and feedback. Fazl also thanks Evan Hubinger and Ryan Kidd for their funding support.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Dario Amodei, Christopher Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. Concrete problems in ai safety. *ArXiv*, abs/1606.06565, 2016. URL https://api.semanticscholar.org/CorpusID:10242377.

Arditi Andy. Unlearning via rmu is mostly shallow. In *Less Wrong*.

Anonymous. CASE-bench: Context-aware safety evaluation benchmark for large language models. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=y9tQNJ2n1y. under review.

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Rishabh Bhardwaj and Soujanya Poria. Language model unalignment: Parametric red-teaming to expose hidden harms and biases. *arXiv preprint arXiv:2310.14303*, 2023.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. Digital forgetting in large language models: A survey of unlearning methods. *arXiv preprint arXiv:2404.02062*, 2024.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

California State Legislature. California Consumer Privacy Act of 2018, 2018. URL https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5. Cal. Civ. Code § 1798.100 et seq.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*, 2023.

Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024a. Curran Associates Inc.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024b.

Zachary Charles, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Nicole Mitchell, Krishna Pillutla, and Keith Rush. Fine-tuning large language models with user-level differential privacy. *arXiv preprint arXiv:2407.07737*, 2024.

Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.738. URL https://aclanthology.org/2023.emnlp-main.738.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled reward mitigates hacking in RLHF. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=zcIV8OQFVF.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911, 2021.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023a. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25879. URL https://doi.org/10.1609/aaai.v37i6.25879.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023b.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581.

Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, and Josivania Silva Farias. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3): 505–525, 2021.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. *arXiv preprint arXiv:2404.12699*, 2024.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=av0D19pSkU.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024b.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 30039–30069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5fc47800ee5b30b8777fdd30abcaaf3b-Paper-Conference.pdf.

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

European Parliament and Council of the European Union. General Data Protection Regulation (GDPR), 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. Regulation (EU) 2016/679.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*, 2024.

Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16352–16362, 2023.

Alex Foote, Neel Nanda, Esben Kran, Ioannis Konstas, and Fazl Barez. N2g: A SCALABLE APPROACH FOR QUANTIFYING INTERPRETABLE NEURON REPRESENTATION IN LLMS. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. URL https://openreview.net/forum?id=ZB6bK6MTYq.

Jasmina Gajcin, James McCarthy, Rahul Nair, Radu Marinescu, Elizabeth Daly, and Ivana Dusparic. Iterative reward shaping using human feedback for correcting reward misspecification. *arXiv preprint arXiv:2308.15969*, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.

Shashwat Goel, Ameya Prabhu, Philip Torr, Ponnurangam Kumaraguru, and Amartya Sanyal. Corrective machine unlearning. *arXiv preprint arXiv:2402.14015*, 2024.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

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

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*, 2024.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Oc1SIKxwdV.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*, 2024.

Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–296, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Yihuai Hong, Lei Yu, Shauli Ravfogel, Haiqin Yang, and Mor Geva. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*, 2024.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Jogging the memory of unlearned model through targeted relearning attack. *arXiv preprint arXiv:2406.13356*, 2024.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *ArXiv*, abs/2404.11045, 2024a. URL https://api.semanticscholar.org/CorpusID:269187650.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*, 2024b.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*, 2024c.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. *arXiv preprint arXiv:2408.06223*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

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

Masaru Isonuma and Ivan Titov. Unlearning traces the influential training data of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6312–6325, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.343.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805.

Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. Sok: Challenges and opportunities in federated unlearning. *arXiv preprint arXiv:2403.02437*, 2024.

Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, and Yaodong Yang. Language models resist alignment. *arXiv preprint arXiv:2406.06144*, 2024.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=VzOgnDJMgh.

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

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. Rwku: Benchmarking real-world knowledge unlearning for large language models, 2024.

Aly Kassem, Omar Mahmoud, and Sherif Saad. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.265. URL https://aclanthology.org/2023.emnlp-main.265.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1885–1894. JMLR.org, 2017.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. 4 2020. URL https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Dmitrii Krasheninnikov, Egor Krasheninnikov, Bruno Kacper Mlodozeniec, Tegan Maharaj, and David Krueger. Implicit meta-learning may lead language models to trust more reliable sources. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Fzp1DRzCIN.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. *arXiv preprint arXiv:2407.04694*, 2024.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=xlr6AUDuJz.

Wenjie Li, Jiawei Li, Christian Schroeder de Witt, Ameya Prabhu, and Amartya Sanyal. Delta-influence: Unlearning poisons via influence functions. *arXiv preprint arXiv:2411.13731*, 2024b.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation, 2023.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022a.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024a.

Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 280–289. IEEE, 2022b.

Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and Xiangyu Zhang. Complex backdoor detection by symmetric feature differencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15003–15013, 2022c.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 1817–1829, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.107.

Ziyao Liu, Huanyi Ye, Chen Chen, and Kwok-Yan Lam. Threats, attacks, and defenses in machine unlearning: A survey. *arXiv preprint arXiv:2403.13682*, 2024c.

Michelle Lo, Shay Cohen, and Fazl Barez. Large language models relearn removed concepts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8306–8323, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.492.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024.

Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3XnBVK9sD6.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. Language model inversion. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=t9dWHpGkPj.

Andrei Muresanu, Anvith Thudi, Michael R Zhang, and Nicolas Papernot. Unlearnable algorithms for in-context learning. *arXiv preprint arXiv:2402.00751*, 2024.

Neel Nanda, Senthooran Rajamanoharan, J'anos Kram'ar, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, Dec 2023. URL https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective, 2024. URL https://arxiv.org/abs/2209.00626.

Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2023.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5716–5731, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.310.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2024. URL https://openreview.net/forum?id=5LhYYajlqV.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024a.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, carsten maple, Subhabrata Majumdar, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=eP9auEJqFg.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2859–2873, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.188. URL https://aclanthology.org/2023.findings-emnlp.188.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Casper Stephen. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=zWqr3MQuNs.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024b.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai. *arXiv preprint arXiv:2407.00106*, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Nexhi Sula, Abhinav Kumar, Jie Hou, Han Wang, and Reza Tourani. Silver linings in the shadows: Harnessing membership inference for machine unlearning. *arXiv preprint arXiv:2407.00866*, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Zou Andy, Song Dawn, Li Bo, Hendrycks Dan, and Mazeika Mantas. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*, 2024.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. I'm not familiar with the name harry potter: Prompting baselines for unlearning in LLMs. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=eBcVsC4h6A.

UK AI Safety Institute. Advanced ai evaluations at aisi: May update, 2024. URL https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

UKAISI. UK AI Safety Institute. https://www.aisi.gov.uk. Accessed: 2024-09-01.

USAISI. US AI Safety Institute. https://www.nist.gov/aisi. Accessed: 2024-09-01.

Àlex Pujol Vidal, Anders S Johansen, Mohammad NS Jahromi, Sergio Escalera, Kamal Nasrollahi, and Thomas B Moeslund. Verifying machine unlearning with explainable ai. *arXiv preprint arXiv:2411.13332*, 2024.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. KGA: A general machine unlearning framework based on knowledge gap alignment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13264–13276, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.740. URL https://aclanthology.org/2023.acl-long.740.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=VJMYOfJVC2.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. Large scale knowledge washing. *arXiv preprint arXiv:2405.16720*, 2024b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Li Tianle, Ku Max, Wang Kai, Zhuang Alex, Fan Rongqi, Yue Xiang, and Chen Wenhu. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024c.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023b.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jA235JGM09.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Sam Boman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. *arXiv:2409.12822*, 2024a.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=84n3UwkH7b.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback, 2024. URL https://arxiv.org/abs/2411.02306.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174.

Sophie Xhonneux, David Dobre, Jian Tang, Gauthier Gidel, and Dhanya Sridhar. In-context learning can re-learn forbidden tasks. *arXiv preprint arXiv:2402.05723*, 2024.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*, 2023.

Heng Xu, Tianqing Zhu, Lefeng Zhang, and Wanlei Zhou. Really unlearned? verifying machine unlearning via influential sample pairs. *arXiv preprint arXiv:2406.10953*, 2024a.

Weiwei Xu, Kai Gao, Hao He, and Minghui Zhou. A first look at license compliance capability of llms in code generation. *arXiv e-prints*, pages arXiv–2408, 2024b.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.457.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024b.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Socially Responsible Language Modelling Research*, 2023. URL https://openreview.net/forum?id=wKe6jE065x.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Yekhanin Sergey, and Zhang Huishuai. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. *arXiv preprint arXiv:2408.10682*, 2024.

Yi Zeng, Weiyu Sun, Tran Ngoc Huynh, Dawn Song, Bo Li, and Ruoxi Jia. Beear: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. *arXiv preprint arXiv:2406.17092*, 2024.

Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=OkChMnjF6s.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations. In *Advances in Neural Information Processing Systems*, 2023a.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *ArXiv*, abs/2312.15710, 2023b. URL https://api.semanticscholar.org/CorpusID:266551298.

Chenxu Zhao, Wei Qian, Yangyi Li, Aobo Chen, and Mengdi Huai. Rethinking adversarial robustness in the context of the right to be forgotten. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60927–60939. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24k.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:267751264.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

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