An Adversarial Perspective on Machine Unlearning for AI Safety

Jakub Łucki[1] **Boyi Wei**[2] **Yangsibo Huang**[2]
Peter Henderson[2] **Florian Tramèr**[1] **Javier Rando**[1]
[1]ETH Zurich [2]Princeton University

Abstract

Large language models are finetuned to refuse questions about hazardous knowledge, but these protections can often be bypassed. Unlearning methods aim at completely removing hazardous capabilities from models and make them inaccessible to adversaries. This work challenges the fundamental differences between unlearning and traditional safety post-training from an adversarial perspective. We demonstrate that existing jailbreak methods, previously reported as ineffective against unlearning, can be successful when applied carefully. Furthermore, we develop a variety of adaptive methods that recover most supposedly unlearned capabilities. For instance, we show that finetuning on 10 unrelated examples or removing specific directions in the activation space can recover most hazardous capabilities for models edited with RMU, a state-of-the-art unlearning method. Our findings challenge the robustness of current unlearning approaches and question their advantages over safety training.[1]

1 Introduction

Large language models (LLMs) are pretrained on trillions of tokens crawled from the Internet (Dubey et al., 2024). Due to the unprecedented size of the training corpora, it is nearly impossible to discard all dangerous or otherwise harmful information available online. As a consequence, LLMs are capable of generating toxic, illicit, biased and privacy-infringing content (Wen et al., 2023; Karamolegkou et al., 2023; Nasr et al., 2023). Since models are constantly becoming more capable, this knowledge may pose increasing risks as it can make hazardous information more easily accessible for adversaries.

LLMs often undergo safety finetuning to reject unethical requests and produce safe responses (Bai et al., 2022). Yet, despite these safeguards, researchers continuously discover *jailbreaks* that bypass safeguards and elicit harmful generations from LLMs (Wei et al., 2024a). Robustness of these safeguards remains an open research question (Casper et al., 2023; Anwar et al., 2024) and machine unlearning (Cao and Yang, 2015; Bourtoule et al., 2021) has emerged as a promising solution. It aims to completely remove hazardous knowledge from LLMs, preventing its extraction even after jailbreaking. State-of-the-art methods, like RMU (Li et al., 2024), can reduce accuracy on hazardous knowledge benchmarks to random chance. However, unlearning is not foolproof, as hazardous knowledge can still be recovered after the process (Patil et al., 2024; Shumailov et al., 2024; Hu et al., 2024). This raises an important question: Does unlearning truly remove hazardous knowledge, or does it simply "obfuscate" this knowledge similarly to refusal safety training?

In this work, we challenge the fundamental differences between unlearning and traditional safety finetuning from an adversarial perspective. We use the accuracy on the WMDP benchmark (Li et al., 2024) to measure the hazardous knowledge contained in LLMs. We argue that, from the perspective

[1]Code is available at: https://github.com/ethz-spylab/unlearning-vs-safety

Preprint.

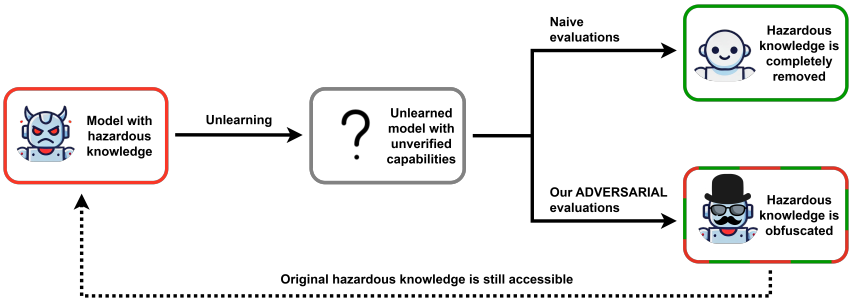

Figure 1: Conceptual overview of our contribution. Our adversarial evaluations show that current unlearning methods largely obfuscate hazardous knowledge instead of erasing it from model weights.

of model safety, unlearning is not successful if there exists *at least one* way of recovering significant accuracy either *without updating the model weights* or *updating the model weights with data that has little or no mutual information with the target knowledge*.

We perform the first comprehensive white-box evaluation of state-of-the-art unlearning methods for hazardous knowledge, comparing them to traditional safety training with DPO (Rafailov et al., 2024). Our results show that while unlearning is robust against specific attacks like probing internal model activations, it can also be easily compromised with methods similar to those used against safety training. Jailbreak methods that were reported ineffective against unlearning, like GCG (Zou et al., 2023), can recover substantial accuracy after small changes in the loss function. Additionally, we find that removing specific directions in the activation space, or finetuning on 10 unrelated examples can completely undo unlearning and recover the original performance on WMDP.

Overall, our findings underscore the limitations of black-box evaluations in accurately assessing unlearning effectiveness for safety settings and highlight the pressing need to refine unlearning methods, so that they deliver their promised benefits over standard safety training.

2 Related Work

Safety training and jailbreaks. Large language models are finetuned to refuse questions about hazardous knowledge with safety methods like DPO (Rafailov et al., 2024) or PPO (Ouyang et al., 2022). Zou et al. (2024) recently introduced *circuit breakers* that use representation engineering to orthogonalize directions corresponding to unwanted concepts. The robustness of existing safeguards is limited (Casper et al., 2023; Anwar et al., 2024) and researchers often find *jailbreaks* to bypass protections and elicit hazardous knowledge (Wei et al., 2024a). Jailbreaks can rely only on prompting strategies (Shah et al., 2023; Huang et al., 2023), exploit white-box access to optimize prompts (Zou et al., 2023; Andriushchenko et al., 2024) or ablate model activations (Arditi et al., 2024).

Unlearning. Unlearning aims to update the weights of a model to remove specific knowledge so that it cannot be accessed in any form (Cao and Yang, 2015; Bourtoule et al., 2021). In the context of language models, unlearning work has expanded across topics like fairness, privacy, safety or hallucinations (Jang et al., 2022; Yao et al., 2023; Chen and Yang, 2023; Wu et al., 2023; Li et al., 2024; Liu et al., 2024b). Unlearning is usually evaluated using narrow topics (e.g. Harry Potter) or fictional information (Eldan and Russinovich, 2023; Maini et al., 2024; Shi et al., 2024; Wei et al., 2024c). Our work focuses on unlearning methods for safety. These methods try to eliminate dangerous knowledge to prevent adversaries from accessing it, even after jailbreaking attempts. The most notable method for this purpose is RMU (Li et al., 2024), which was introduced alongside WDMP, a benchmark for evaluating hazardous capabilities. General-purpose unlearning algorithms like negative preference optimization (NPO) (Zhang et al., 2024) can also be adapted for this purpose.

Unlearning robustness. Initial unlearning evaluations for LLMs relied on simple classification metrics (Eldan and Russinovich, 2023) which do not account for all possible ways in which a language model can represent and output the target information. Recent works (Jin et al., 2024; Hong et al., 2024; Lynch et al., 2024; Schwinn et al., 2024; Pawelczyk et al., 2024) have adopted an adversarial approach to test whether there exist ways to extract the information that was supposedly unlearned.

For instance, Lynch et al. (2024) showed that knowledge could be extracted at comparable rates from both original and unlearned models by probing internal representations. In the context of unlearning hazardous capabilities, RMU reports robustness under some white-box jailbreaks like GCG or probing, but finds that finetuning unlearned models can easily disable the protections (Li et al., 2024). Similarly, Hu et al. (2024) find that fine-tuning can revert unlearning. In this work, we devise novel methods to extract hazardous knowledge from unlearned models without updating the weights. The importance of meticulous evaluations, has been demonstrated by an earlier work on word embedding debiasing, which revealed the lack of robustness of the respective methods (Gonen and Goldberg, 2019).

3 Experimental Setup

This works focuses exclusively on unlearning methods for safety that remove hazardous knowledge (e.g. bioweapons) from large language models, as introduced by Li et al. (2024). In practice, unlearning relies on *forget* and *retain* sets. The first contains information relevant to the domain to be unlearned (e.g. enhanced pandemic pathogens) while the second includes neighboring information that should be preserved (e.g. general biology). In this work, we use the datasets included in WMDP benchmark for biology and cybersecurity (Li et al., 2024). Our evaluation is designed to assess whether existing unlearning methods effectively remove hazardous knowledge or merely make it more difficult to access, similarly to safety training.

3.1 Threat Model

We assume white-box access to an unlearned model, allowing modification of its weights and intervention in the activation space during inference. Additionally, we assume access to the original model prior to unlearning or to an equivalent model obtained by removing unlearning protections through finetuning, as demonstrated later. Although white-box access differs from the threat model for protections we study (RMU assumes only black-box access), it provides valuable insights into the effectiveness of unlearning in removing knowledge from model weights. Furthermore, with the rise of powerful open-source large language models, robust unlearning in white-box scenarios is an increasingly relevant desiderata.

3.2 Unlearning Methods and Safety Training Baseline

We evaluate the most powerful unlearning method for hazardous knowledge to date: RMU (Li et al., 2024; Kadhe et al., 2024)[2]. Additionally, we implement NPO (Zhang et al., 2024) that has been widely used as a general-purpose unlearning method for fact and concept removal (Shi et al., 2024), but its effectiveness for hazardous knowledge removal remains unexplored. We specifically use NPO+RT, a variant of NPO including an additional retain loss. Finally, we include DPO (Rafailov et al., 2024) as a baseline for safety training to contrast it with unlearning methods. For more details about the methods, see Appendix B.

3.3 Models and Datasets

We evaluate the performance of RMU using the publicly available checkpoint[3]. This model results from finetuning Zephyr-7B-β (Tunstall et al., 2023) on the WMDP and WikiText corpora (Merity et al., 2016). For NPO and DPO, we finetune Zephyr-7B-β ourselves on WMDP. We will refer to these models as *unlearned models*.

NPO and DPO require preference datasets, but WMDP only provides corpora (e.g. scientific papers) for autoregressive training. We use GPT-4 (OpenAI et al., 2024) to formulate questions based on these documents. For questions about hazardous topics, we set one of 80 random refusal strings as the desired output and the full correct option as the rejected response. For questions based on the

[2] *Embedding-COrrupted Prompts* (ECO) (Liu et al., 2024a) outperforms others but applies a pre-LLM filter, leaving the original weights and potential hazardous knowledge unchanged. Thus, it doesn't meet our definition of unlearning. See Appendix A for further discussion.

[3] Available at `https://huggingface.co/cais/Zephyr_RMU`

retain set, we keep the correct option as the desired output and reject the refusal. We refer to the resulting datasets as our *preference datasets*. See Appendix C for details on dataset construction.

To ensure a fair comparison with safety methods, we fine-tune Zephyr using DPO specifically on preference datasets relevant to unlearning topics, rather than training it to refuse all harmful requests. We balance the training data by including samples from the forget and retain preference datasets, as well as OpenAssistant (Köpf et al., 2024), in a 50:25:25 ratio. This approach aims to maintain a balance between refusal capabilities and preserving general utility. For NPO, we use the preference dataset on hazardous knowledge as negative samples and the retain preference dataset mixed with OpenAssistant (50:50) dataset for the auxiliary retain loss.

3.4 Unlearning Evaluation

We evaluate the performance of unlearning hazardous knowledge using the WMDP benchmark (Li et al., 2024), which consists of 1,273 multiple-choice questions about dangerous biology knowledge and 1,987 about cybersecurity. To detect latent knowledge that might still be present even when models refuse to answer, we select the option (A, B, C, or D) with the highest probability as the final response. Besides, we use MMLU (Hendrycks et al., 2020) to measure the model's general utility after unlearning, which contains multiple-choice questions covering 57 different tasks. For both WMDP benchmark and MMLU, we report overall accuracy across the entire dataset.

4 Our Methods To Recover Hazardous Capabilities

We use a wide range of methods to uncover hazardous capabilities in the target models, ranging from representation engineering to prompt-based jailbreaks. Most methods are inspired by well-known safety jailbreaks and incorporate small changes to target unlearning methods. All of our methods—except for finetuning—do not modify model weights and, thus, can only access knowledge that was preserved in model weights after unlearning. For finetuning, we primarily use small or unrelated datasets to ensure that models cannot acquire new hazardous capabilities.

4.1 Finetuning

It has been shown that finetuning easily reverses safety alignment even when using benign datasets (Qi et al., 2023). Also, the original RMU work and showed that fine-tuning unlearned models on the entire forget dataset could recover hazardous capabilities. In this work, we fine-tune unlearned models on datasets with very low mutual information (MI) with the unlearned knowledge to ensure that no new knowledge can be acquired. We use Low-Rank Adaptation (LoRA; Hu et al., 2021) to fine-tune unlearned models on three datasets: (1) forget dataset, (2) retain dataset—disjoint with forget dataset by definition—, and (3) WikiText (Merity et al., 2016)—a collection of Wikipedia documents with minimal overlap with hazardous knowledge. We experiment with varying sample sizes (from 5 to 1000 examples). By incorporating datasets with high MI (forget set) and low MI (retain set and WikiText), we provide a comprehensive evaluation of how different configurations affect the pace of hazardous knowledge recovery. For further details see Appendix E.1.

4.2 Orthogonalization

Arditi et al. (2024) demonstrated that safety refusal is governed by a single direction in the activation space. We investigate whether unlearning techniques generate a similar direction. Rather than targeting a single layer, we allow for distinct refusal directions at each transformer block. Using the forget preference dataset, we collect the outputs of each transformer block from both the original and unlearned models. We then compute the refusal direction for each layer using the difference in means method (Belrose, 2023). At inference time, we remove the refusal direction at each layer. Additionally, we develop a setup that does not require access to the original model prior to unlearning; see Appendix E.2 for details.

4.3 Logit Lens

Logit Lens is an interpretability technique (nostalgebraist, 2020; Patil et al., 2024) that projects the activations in residual stream onto the model's vocabulary. We apply this technique to the WMDP

Table 1: WMDP-Bio and MMLU accuracy for each protection and method. For Logit Lens, we report the best layer overall. For finetuning, we report best result on 5 samples from the forget set. Empty values are not possible to compute or the corresponding combination does not affect the score.

Datasets	Knowledge Recovery	No Protection	Unlearning Methods		Safety Training
			RMU	NPO	DPO
WMDP-Bio	Default decoding	64.4	29.9	29.5	27.9
	Logit Lens	66.2	31.8	38.6	48.2
	Finetuning	-	62.4	47.4	57.3
	Orthogonalization	-	64.7	45.1	50.7
	Enhanced GCG	-	53.9	46.0	49.0
	Pruning	-	54.0	40.4	50.4
MMLU	Default decoding	58.1	57.1	52.1	49.7
	Logit Lens	-	-	-	-
	Finetuning	-	58.0	53.3	51.2
	Orthogonalization	-	57.3	45.6	46.7
	Enhanced GCG	-	-	-	-
	Pruning	-	56.5	50.0	50.4

dataset by using the projected logits of the A, B, C, and D tokens as the model's answers. We project the output of transformer blocks at every layer and select the token with a higher probability. We also evaluate the projection of other activation spaces in Appendix G.3.

4.4 Enhanced GCG

GCG has been reported ineffective against RMU (Li et al., 2024; Huu-Tien et al., 2024). We introduce *enhanced GCG*, which especifically targets unlearning methods, and is based on FLRT (Thompson and Sklar, 2024) and augmented with several modifications detailed in Appendix E.4. Unlike GCG, which aims to find adversarial prompt suffixes, enhanced GCG focuses on optimizing *prefixes* to prevent the model from recognizing hazardous knowledge in the first place, as RMU will introduce persistent noise to the residual stream once such context is detected. We also attribute more weight to the loss computed on early tokens in the prompt. Our attack is optimized on 6 questions from the WMDP benchmark that were answered correctly by the original model and incorrectly by the unlearned model.

4.5 Set difference pruning

Wei et al. (2024b) introduced *set difference pruning* as a method to identify and prune neurons associated with safety alignment. Reproducing their method, we use SNIP (Lee et al., 2018) score to measure the importance of individual neurons for hazardous knowledge. Specifically, we compute the importance score for each neuron on the WMDP forget set, and the utility score on MMLU. We then use set difference method to find the neurons that only contribute to storing hazardous knowledge and remove them via pruning.

5 Results

We report the performance of our methods on WMDP-Bio due to significant difference in the scores of original and unlearned models. Analogous gap on WMDP-Cyber is much smaller, which makes the corresponding results more volatile (See Appendix F). We summarize our results and observations below.

Finetuning on unrelated information reverts unlearning. As illustrated in Figure 2, finetuning with only 10 samples from the retain set—disjoint by definition from the evaluation knowledge—can recover most of hazardous capabilities, obtaining accuracies of 52.7% (NPO), 57.0% (DPO), and 61.6% (RMU) while causing negligible degradation on MMLU (less than 2 p.p.). Finetuning on

1000 samples from the retain set fully recovers hazardous capabilities across all methods. These results demonstrate that both safety training and unlearning can be undone through finetuning on unrelated information, suggesting that unlearning is also expressed through shallow features (Yang et al., 2023; Lermen et al., 2023). Additionally, finetuning with just 5 samples from the forget set effectively reverses unlearning, particularly for RMU, which nearly recovers its original performance. Relearning knowledge through further training is unavoidable, but these results show that knowledge recovery happens at disproportionately fast rate.

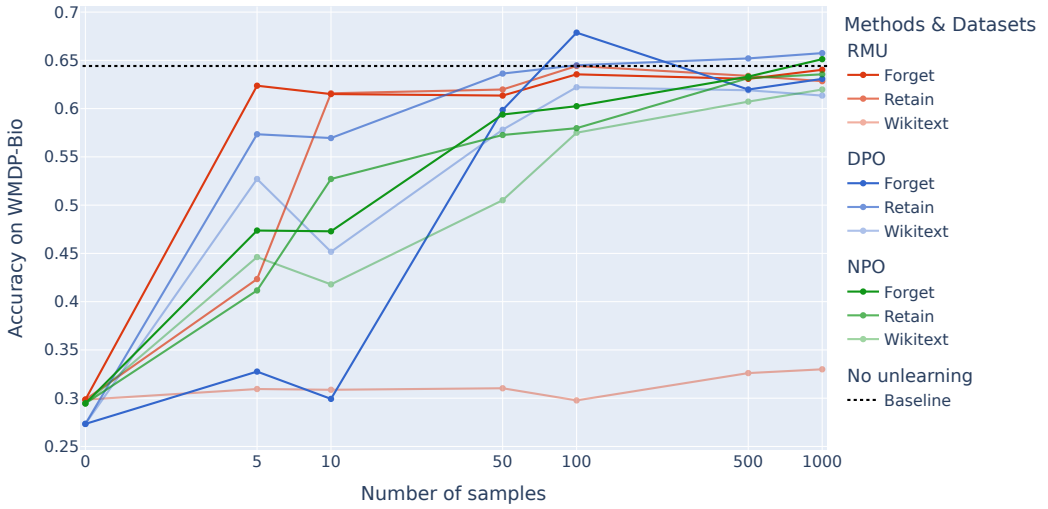

Figure 2: Accuracy on WMDP-Bio for unlearned models finetuned with different datasets and number of samples. See Appendix F.1 for complimentary results on MMLU and WMDP-Cyber.

Unlearning methods remove knowledge from the residual stream more effectively. Before unlearning, Logit Lens can decode correct answers from Zephyr-7B at layer 19, as shown in Figure 3. However, Logit Lens becomes ineffective after protections are applied. Our safety baseline, DPO, remains the most susceptible to early decoding, achieving 56% accuracy. In contrast, unlearning methods can remove knowledge more effecitvely from the residual stream, with RMU reducing Logit Lens accuracy close to random chance across the entire architecture. These results align with prior evaluations of RMU's robustness to probing (Li et al., 2024).

Unlearning is also mediated by specific directions. We identify and ablate directions responsible for unlearning, successfully recovering hazardous knowledge for most protections (see Table 1). RMU is the most vulnerable to our orthogonalization, achieving 64.7% accuracy (surpassing the

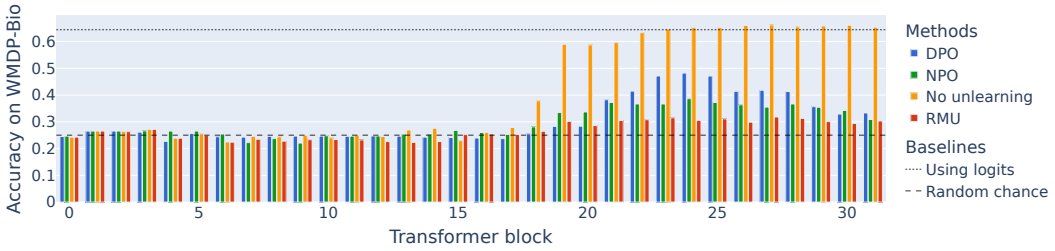

Figure 3: Accuracy on WMDP-Bio using LogitLens after each transformer block.

baseline accuracy of 64.4%) by manipulating only the activation space during the forward pass. This outperforms ablation of a single refusal direction across all layers (Arditi and Chughtai, 2024), which achieves 54.2% accuracy. NPO and DPO are more robust against orthogonalization, obtaining 45% and 51% accuracy, respectively.

Unlearning depends on critical neurons. We localized minimal sets of weights that are responsible for degradation in hazardous knowledge for each unlearning method. These sets represent 2.0% of weights for NPO, 0.9% for RMU, and 2.4% for DPO. After pruning these weights, performance on WMDP increases by at least 10 p.p. for all methods.

Universal adversarial prefixes that recover unlearned knowledge exist. Using *enhanced GCG* we were able to craft universal adversarial prefixes that increased RMU's accuracy from 29.9% to 53.9%, NPO's accuracy from 29.5% to 46.0%, and DPO's accuracy from 27.9% to 49.0%. This demonstrates that, similarly to safety trained models, input-only manipulations can disable unlearning and elicit hazardous knowledge that was never removed from the model.

We can recover hazardous capabilities while models remain unusable. RMU is characterized by making models unusable—they output gibberish generations with high perplexity—when hazardous knowledge is detected. Interestingly, we find that GCG prefixes can easily recover a conversational model that answers questions from WMDP, but its responses are often incorrect and overconfident. Best performing prefixes can recover most of the hazardous capabilities while not necessarily recovering conversational capabilities from the model. See Appendix I for an analysis.

6 Discussion

Existing unlearning methods are not different from safety training. Our findings reveal that unlearning methods primarily obscure knowledge rather than eliminate it (as illustrated by Figure 1), which is a known flaw of safety training (Lee et al., 2024). Therefore, RMU and NPO are susceptible to techniques analogous to those that can reverse safety training, including: (1) dependence on individual residual stream directions; (2) rapid knowledge recovery after finetuning with unrelated data; (3) presence of critical neurons that inhibit hazardous knowledge; and (4) existence of universal adversarial strings that unlock the unlearned knowledge. These observations question the practical benefits of unlearning methods over safety training. Although unlearning was proposed to fully eradicate hazardous capabilities and mitigate jailbreaks in large language models, our results indicate that these methods share limitations. Concurrent work by Tamirisa et al. (2024) proposed TAR, a technique that can prevent *some* fine-tuning attacks but has no impact on others.

Black-box evaluations are insufficient for unlearning. Our results demonstrate that evaluations based solely on model outputs are not suitable for unlearning methods, as suggested previously by Lynch et al. (2024). Unlearning aims to remove information from model weights, which is fundamentally different from merely rendering knowledge unusable in downstream tasks. We thus argue that output-based evaluations are insufficient and future evaluations should prioritize measuring the extent to which knowledge is genuinely erased from the model weights. As extensively demonstrated in security and safety research, adaptive evaluations are important to understand the limitations of ML protections (Carlini and Wagner, 2017; Tramer et al., 2020; Radiya-Dixit et al., 2021; Hönig et al., 2024)

NPO shows signs of deep unlearning. This method consistently displays better robustness than DPO or RMU, suggesting that gradient ascent (Zhang et al., 2024) might be a promising tool to remove hazardous knowledge from model weights. However, our current implementation still results in greater degradation on MMLU and general capabilities. Future work could investigate combining representation engineering with gradient ascent to enhance existing unlearning methods.

7 Conclusion

We performed a comprehensive white-box evaluation of state-of-the-art unlearning methods for AI safety. Our findings reveal that these methods cannot reliably remove knowledge from model weights.

For example, finetuning on unrelated data or removing specific directions from actiavtion space often recovers the supposedly unlearned capabilities. This challenges the belief that unlearning methods offer more robust protection than standard safety training. Furthermore, we argue that black-box evaluations are insufficient for unlearning, as they do not assess internal model changes.

Acknowledgement

Thanks to Daniel Paleka, and Stephen Casper for useful discussions. Research reported in this publication was supported by an Amazon Research Award Fall 2023. JR is supported by an ETH AI Center Doctoral Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Amazon.

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks, 2024. URL https://arxiv.org/abs/2404.02151.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Survey Certification, Expert Certification.

Andy Arditi and Bilal Chughtai, Jul 2024. URL https://www.lesswrong.com/posts/6QYpXEscd8GuE7BgW/unlearning-via-rmu-is-mostly-shallow.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Nora Belrose. Diff-in-means concept editing is worst-case optimal: Explaining a result by Sam Marks and Max Tegmark, 2023. https://blog.eleuther.ai/diff-in-means/. Accessed on: September 12, 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.

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.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Survey Certification.

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

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, et al. llama3family, 2024. URL https://arxiv.org/abs/2407.21783.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023. URL https://arxiv.org/abs/2310.02238.

Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. Practical unlearning for large language models. *arXiv preprint arXiv:2407.10223*, 2024.

Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based adversarial examples for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.498. URL https://aclanthology.org/2020.emnlp-main.498.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019. URL https://arxiv.org/abs/1903.03862.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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.

Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

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.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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. *arXiv preprint arXiv:2406.10890*, 2024.

Swanand Ravindra Kadhe, Farhan Ahmed, Dennis Wei, Nathalie Baracaldo, and Inkit Padhi. Split, unlearn, merge: Leveraging data attributes for more effective unlearning in llms. *arXiv preprint arXiv:2406.11780*, 2024.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models, 2023. URL https://arxiv.org/abs/2310.13771.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity, 2024. URL https://arxiv.org/abs/2401.01967.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

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

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*, 2024a.

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*, 2024b.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms, 2024. URL https://arxiv.org/abs/2402.16835.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023. URL https://arxiv.org/abs/2311.17035.

nostalgebraist. Interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

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

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7erlRDoaV8.

Martin Pawelczyk, Jimmy Z Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216*, 2024.

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.

Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramèr. Data poisoning won't save you from facial recognition. *arXiv preprint arXiv:2106.14851*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Vinu Sankar Sadasivan, Shoumik Saha, Gaurang Sriramanan, Priyatham Kattakinda, Atoosa Chegini, and Soheil Feizi. Fast adversarial attacks on language models in one gpu minute. *arXiv preprint arXiv:2402.15570*, 2024.

Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Gunnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *arXiv preprint arXiv:2402.09063*, 2024.

Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. Scalable and transferable black-box jailbreaks for language models via persona modulation, 2023. URL https://arxiv.org/abs/2311.03348.

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*, 2024.

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.

Antoine Simoulin and Benoit Crabbé. How many layers and why? An analysis of the model depth in transformers. In Jad Kabbara, Haitao Lin, Amandalynne Paullada, and Jannis Vamvas, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 221–228, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-srw.23. URL https://aclanthology.org/2021.acl-srw.23.

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

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

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. URL https://arxiv.org/abs/2310.16944.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024a.

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*, 2024b.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating copyright takedown methods for language models. *arXiv preprint arXiv:2406.18664*, 2024c.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models, 2023. URL https://arxiv.org/abs/2311.17391.

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. *arXiv preprint arXiv:2310.20138*, 2023.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

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

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

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL https://arxiv.org/abs/2406.04313.