ULMR: Unlearning Large Language Models via Negative Response and Model Parameter Average

Shaojie Shi[*♠] Xiaoyu Tan[*♡] Xihe Qiu[*◇†] Chao Qu[♡] Kexin Nie[♣]
Yuan Cheng[♠] Wei Chu[♡] Yinghui Xu[♠] Yuan Qi[♠]
[♡] INF Technology (Shanghai) Co., Ltd. [◇] Shanghai University of Engineering Science
[♠] AI^3 Institute, Fudan University [♣] Sunrise Life Network Technology Co., Ltd.
yulin.txy@inftech.ai, qiuxihe1993@gmail.com

Abstract

In recent years, large language models (LLMs) have attracted significant interest from the research community due to their broad applicability in many language-oriented tasks, and are now widely used in numerous areas of production and daily life. One source of the powerful ability of LLMs is the massive scale of their pre-training dataset. However, these pre-training datasets contain many outdated, harmful, and personally sensitive information, which inevitably becomes memorized by LLM during the pre-training process. Eliminating this undesirable data is crucial for ensuring the model's safety and enhancing the user experience. However, the cost of extensively cleaning the pre-training dataset and retraining the model from scratch is very high. In this work, we propose ULMR , an unlearning framework for LLMs , which first uses carefully designed prompts to rewrite the instructions in the specified dataset, and generate corresponding negative responses. Subsequently, to ensure that the model does not excessively deviate post-training, we perform model parameter averaging to preserve the performance of the original LLM. We conducted experiments on two public datasets, TOFU and RWKU, demonstrating that our method can effectively forget specified information while retaining the capabilities of the original LLM.

1 Introduction

Large language models (LLMs) have achieved commendable success in various tasks, demonstrating their capability to disseminate knowledge across different fields and tasks. Nowadays, LLMs are being utilized by the general public as personal assistants, providing advice and solutions for a variety of daily activities (Perez et al., 2022; Menick et al., 2022; Kadavath et al., 2022; Bai et al., 2022). The remarkable abilities of LLMs largely stem from the massive dataset used during their pre-training process. LLMs can parameterize this knowledge, possessing the ability to recall and apply it when generating responses. However, the pre-training dataset widely contains personal privacy information (such as personal identification codes) and harmful content, including biases, discrimination, or content that violates human ethics. Additionally, using copyrighted content without consent for pre-training has garnered attention. Many countries have privacy protection laws requiring that personal data not be disclosed arbitrarily or allowing individuals or organizations to request the deletion of their data from service providers according to their wishes (Hoofnagle et al., 2019; Pardau, 2018).

A straightforward approach is to inspect the pre-training dataset, remove problematic data, and then retrain the model from scratch using the remaining dataset (Kumar et al., 2022). This method has been widely applied in smaller-scale neural network models, but it is prohibitively expensive and impractical for LLMs with billions of parameters. Therefore, the method of fast approximate unlearning is crucial. Research on unlearning is still in its early stages, focusing on the fields of machine learning, and unlearning for LLMs remains a challenging task (Zhao et al., 2024).

In this paper, we propose a framework named ULMR for rapid and efficient forgetting on specific instruction sets for LLMs. First, we enhance the model's ability to generalize and improve its forgetting performance by rewriting the initial instruction set using carefully designed prompts. Second, based on the rewritten instructions, we generate corresponding negative responses to train the LLM to produce confused responses about the information to be forgotten. Finally, to ensure that the weight shift of the model post-training is controlled, we perform a model parameter averaging process to maintain the model's general capabilities without significant degradation.

[*] Equal Contributions.
[†] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 755–762
November 12-16, 2024 ©2024 Association for Computational Linguistics

Supervised Finetuning (SFT) by providing specific tasks or directives to the model, enables it to better understand and execute different types of tasks and is a vital method for updating the model's knowledge base (Bakker et al., 2022; Lou et al., 2023). SFT is also applied in many LLM unlearning algorithms. The instruction rewriting process can alleviate the overfitting of patterns in the training data by LLMs during training, thereby enhancing their generalization capabilities. Model parameter averaging can mitigate the adverse effects on the model's capabilities during the unlearning process, striking a better balance between forgetting and general capabilities (Wortsman et al., 2022). Our empirical results from experiments demonstrate that the framework we propose can effectively forget knowledge on specified data while maximally preserving its general capabilities.

2 Related Works

2.1 Machine Unlearning

The goal of Machine Unlearning is to eliminate a trained model's memory of a subset of its training data (Nguyen et al., 2022). Initially applied extensively in the field of computer vision for image classification tasks, it was used to make models forget specific image categories to achieve balanced classification performance or protect privacy. A common method involves using the Fisher Information Matrix to measure the sensitivity of model outputs to parameter perturbations, thereby inducing the model to "forget"(Golatkar et al., 2020; Foster et al., 2024). For diffusion generative models, a reverse Teacher-Student model can guide the unlearning process (Gandikota et al., 2023). In fact, Machine Unlearning is a challenging process, influenced by the neural network's memory capabilities and the similarity between the forgetting set and the retain set (Zhao et al., 2024).

By designing special prompts or using In-Context Learning (Pawelczyk et al., 2023) techniques, models can appear to have forgotten the targeted knowledge without additional training, although this method is heavily influenced by the model's inherent performance (Jin et al., 2024). More commonly, methods focus on reducing the impact of adverse data through Supervised finetuning processes, such as Gradient Ascent (Jang et al., 2022) and KL Minimization (Maini et al., 2024). Additionally, Direct Preference Optimization (DPO) (Rafailov et al., 2024) and Negative Preference Optimization (NPO) (Zhang et al., 2024), built on the concept of reinforcement learning, are effective LLM unlearning algorithms. However, studies indicate that even after unlearning, LLMs might "forget" how to apply the forgotten knowledge, but these pieces of knowledge could still potentially exist within the model (Patil et al., 2023).

2.2 LLM Safety

Currently, the internal workings of many LLMs remain opaque, leading to outputs that are complex and difficult to predict. Moreover, the pre-training corpora of these models still contain much harmful information. As the application of LLMs becomes more widespread, concerns about their ethical and security aspects have arisen. The safety of LLMs has thus become a highly prominent topic. Integrating LLMs with human values is a crucial step to ensure their consistent and safe deployment. Askell et al. (2021) have proposed the concept of "HHH", which stands for Helpful, Honest, and Harmless. An exemplary LLM should be helpful to humans, and possess the capabilities of being harmless, protecting privacy, and resisting malicious attacks.

3 Methods

In this work, our goal is to develop a simple and efficient LLM unlearning framework that can forget content in the target dataset while maximizing the retention of general capabilities. Initially, we enhance and restructure the forgotten dataset D_f to maximize the assurance that the model forgets the corresponding knowledge. Subsequently, we perform model parameter averaging to restore the general capabilities of the SFT Model. The complete framework is illustrated in Figure 1. To maximally induce the LLM to forget the content on the specified dataset, we first need to enhance and restructure the forgotten dataset D_f.

3.1 Restructured Dataset

For a given dataset D, we assume the subset D_f that needs to be forgotten is a subset of D. The retained dataset can then be represented as $D_h = D \setminus D_f$. Our goal is to ensure that the model retains inference utility on D_h while forgetting the labeled sequences in D_f.

Firstly, we rewrite the instructions $x_i \in D_f$ using a LLM p_θ through carefully designed $rewrite\_prompt$. This process can be represented

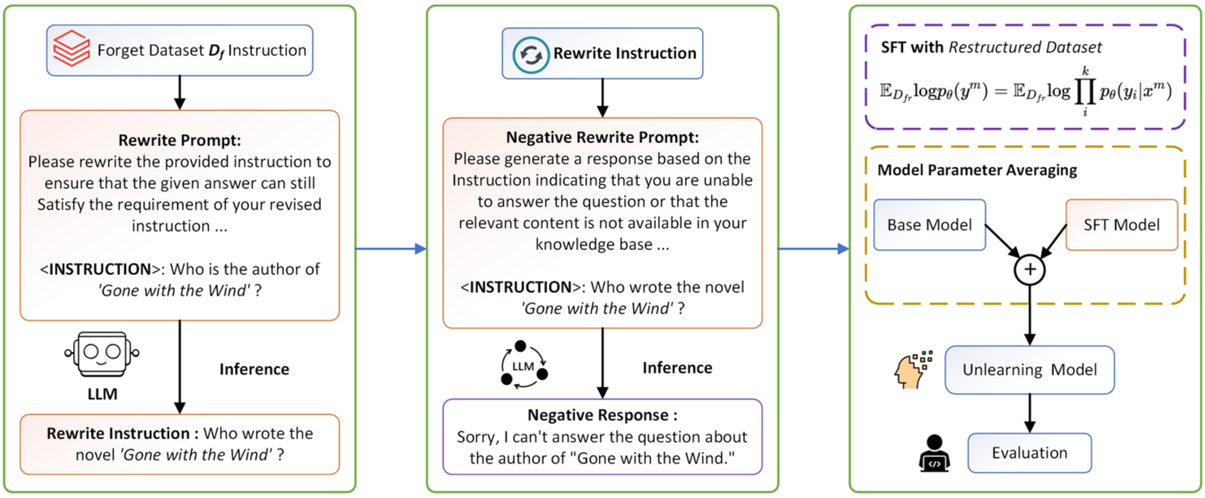
Forget Dataset D_f Instruction
Rewrite Prompt:
Please rewrite the provided instruction to ensure that the given answer can still Satisfy the requirement of your revised instruction ...
<INSTRUCTION>: Who is the author of 'Gone with the Wind' ?
LLM
Inference
Rewrite Instruction : Who wrote the novel 'Gone with the Wind' ?
Rewrite Instruction
Negative Rewrite Prompt:
Please generate a response based on the Instruction indicating that you are unable to answer the question or that the relevant content is not available in your knowledge base ...
<INSTRUCTION>: Who wrote the novel 'Gone with the Wind' ?
LLM
Inference
Negative Response :
Sorry, I can't answer the question about the author of "Gone with the Wind."
SFT with Restructured Dataset
$\mathbb{E}_{D_{fr}}\text{log}p_\theta(y^m) = \mathbb{E}_{D_{fr}}\text{log}\prod_i^k p_\theta(y_i|x^m)$
Model Parameter Averaging
Base Model
SFT Model
Unlearning Model
Evaluation

Figure 1: An overview of ULMR framework

as:

$$x' \sim p_\theta(\cdot \mid rewrite\_prompt(x_i)),$$

The $prompt\_rewrite$ is shown in Table 1.

$prompt\_rewrite$
Please rewrite the provided instruction to ensure that the given answer can still satisfy the requirement of your revised instruction. Instruction:"{instruction}" Rewrite Instruction:

Table 1: The prompt of $prompt\_rewrite$.

We rewrite each instruction x_i twice, thus obtaining a rewritten instruction set $x_r = \{x_1, x_2, x_i\}$. Afterward, using the $negative\_prompt$, the rewritten instruction set x_r, we generate the corresponding negative responses $y_r = \{y_1, y_2, y_i\}$. The $negative\_prompt$ is shown in Table 2. In the negative responses, the model can refuse to answer the question or obscure the main entities from the original answer. Rewriting instructions multiple times can enhance the model's generalization ability, preventing the model from learning only fixed patterns in the dataset, which could lead to poor forgetting effects. Negative responses are crucial for inducing the forgetting phenomenon. The processed data can be used to create an enhanced dataset, represented as $(x, y) \in D_{fr}$. Additionally, to maintain compatibility with previous chat model inputs (denoted as θ), we assume that formatting prompts or special tokens used for formatting are known and have already been appended to the instructions x.

$negative\_prompt$
Please generate a response based on the Instruction indicating that you are unable to answer the question or that the relevant content is not available in your knowledge base. Instruction:"{instruction}" Answer:

Table 2: The prompt of $negative\_prompt$.

3.2 Fine-tuning with negative responses

Here, we execute Supervised Fine-tuning on the reconstructed augmented data instruction set D_{fr} containing M data points with the base model p_θ. Each sample in the instruction set D_{fr} contains a rewritten instruction x^m and a negative response y^m, with many tokens in each data point. Typically, SFT is conducted by maximizing the log-likelihood of the response y^m for the overall instruction sample x^m, which can be represented as:

$$\mathbb{E}_{D_{fr}}\text{log}p_\theta(y^m) = \mathbb{E}_{D_{fr}}\text{log}\prod_i^k p_\theta(y_i|x^m) \tag{1}$$

with i and k tokens on each instruction and response, respectively. The major difference between SFT and autoregressive training in the pre-training phase is that we optimize θ by maximizing the log-likelihood on the conditional probability. After undergoing SFT, we can obtain the new model p_f.

3.3 Model Parameter Average

In the field of deep learning, the technique of Model Weight Averaging is employed to improve the performance and stability of models. Studies have shown that averaging the parameters of a model can address the issue of Catastrophic Forgetting in LLMs during Continual Instruction Finetuning(Lin et al., 2023). This technique also helps in regaining some of the general capabilities that are lost in the process. For a primary model denoted by θ and its fine-tuned version θ', the method of Model Parameter Averaging is mathematically represented as:

$$\theta_a = \alpha\theta + (1 - \alpha)\theta'$$

Here, α is a hyperparameter. We perform the model parameter averaging process on the base model p_θ and the SFT model p_f, resulting in the final unlearning model p_u.

3.4 ULMR

The algorithm of ULMR is shown in Algorithm 1.

Algorithm 1 Algorithm of ULMR

```
Inputs: Forget Dataset $D_f$ which contains instruction $x_i$ and response $y_i$ ; base model $p_\theta$ ; $prompt\_rewrite$ ; $negative\_prompt$
for each step do
    1. Rewrite the instructions $x_i \in D_f$ using a LLM $p_\theta$ through $rewrite\_prompt$, get instruction $x' \sim p_\theta(\cdot \mid rewrite\_prompt(x_i))$
    2. Rewrite each instruction $x_i$ twice, thus obtaining a rewritten instruction set $x_r = \{x_1, x_2, x_i\}$
    3. Using the $negative\_prompt$, the rewritten instruction set $x_r$, and the response $y \in D_h$, generate the corresponding negative responses $y_r = \{y_1, y_2, y_i\}$
    4. Building a Restructured Dataset $D_{fr}$ by $x_r$ and $y_r$
    5. Supervised fine-tuning model $p_\theta$ on dataset $D_{fr}$ to get model $p_f$
    6. Perform Model Parameter Averaging on $p_f$ and $p_\theta$, to obtain $p_u$.
end for
return: The Unlearning Model $p_u$
```

4 Experiment

In this section, we will provide a detailed description of our experiment settings, baseline, and benchmark.

4.1 TOFU Unlearning Benchmark

We first conduct experiments on TOFU (Maini et al., 2024), a benchmark specifically designed to evaluate the unlearning capabilities of LLMs. The TOFU Unlearning Benchmark provides a dataset comprising 200 diversified fictional author profiles, each containing 20 question-answer pairs, with a subset forming the forget set. Since all data is fictional, there is no pre-existing prior knowledge in current LLMs related to it, creating a clean unlearning setting and environment. This setup enables a clear delineation of the information scope required to be forgotten. The TOFU dataset consists of four parts:

- **World Fact** Includes basic common knowledge and information about the real world. After the unlearning process, the model should retain all knowledge related to the real world.
- **Forget Set**: The data that the model needs to forget.
- **Retain Set**: The remaining fictional author knowledge that the model must remember after the unlearning process.
- **Real Author**: Examples containing information about real authors.

The forget set is used to evaluate the quality of the model's unlearning, while the other datasets assess the model's general capability. After the unlearning process, the model's performance on datasets outside the forget set should be close to that of the base model. Moreover, due to the effect of knowledge entanglement, it becomes challenging for the model to remember data highly similar to the forget set.

Following the setup by Maini et al. (2024), we report the following metrics on the TOFU dataset to comprehensively evaluate the efficacy of our proposed unlearning algorithm and the model's general capability:

- **ROUGE** (Lin, 2004): Given that the model's output pattern may slightly differ, we use the ROUGE Score as a substitute for accuracy to assess the similarity between the model's output and the reference answers. A higher ROUGE score indicates closer resemblance to the reference answers.
- **Probability**: Assesses the conditional probability of the correct answer given a prompt.

- **Truth Ratio**: Evaluates the likelihood of generating the correct answers. This metric measures the extent to which the designed unlearning algorithm removes information. The Truth Ratio is calculated as follows:

$$R_{\text{truth}} = \frac{\frac{1}{|\mathcal{A}_{\text{pert}}|} \sum_{\hat{a} \in \mathcal{A}_{\text{pert}}} P(\hat{a} \mid q)^{\frac{1}{|\hat{a}|}}}{P(\tilde{a} \mid q)^{\frac{1}{|\tilde{a}|}}}$$

  Here, $\hat{a}$ represents a paraphrased answer, q is the question, and $\mathcal{A}_{\text{pert}}$ consists of perturbed answers generated by GPT-4 (OpenAI, 2023), maintaining the general form of the answers but factually incorrect.

4.2 RWKU Unlearning Benchmark

Similar to the TOFU dataset, RWKU is an unlearning benchmark designed by Jin et al. (2024) to evaluate the unlearning capabilities of LLMs . However, The slight difference between RWKU and TOFU is that it selects 200 well-known real-world figures as unlearning targets, who are typically included in the pre-training corpora of LLMs. The objective of the unlearning algorithm is to make the LLM forget factual knowledge about these targets without affecting related knowledge and overall capabilities. The RWKU dataset comprises four parts:

- **Forget Set**: Records the data that the model needs to forget.

- **Neighbor Set**: Used to assess the model's performance on data that is closely related to but not entirely contained within the unlearning targets.

- **MIA Set**: Utilized to infer whether the model still retains knowledge about the targets.

- **Utility Set**: Evaluates the model's general capabilities.

4.3 Baseline

Currently, many researchers have proposed various more efficient and practical unlearning algorithms. We selected the most representative algorithms as baselines to evaluate the performance of our proposed ULMR framework.

- **Gradient Ascent** (Jang et al., 2022): One of the most common unlearning algorithms. Unlike the typical gradient descent optimization in neural networks, the objective of Gradient Ascent is to maximize the negative log-likelihood loss on the forget set, steering the model away from its initial predictions and promoting the unlearning process.

- **DPO** (Rafailov et al., 2024): Generally, the DPO algorithm requires both positive and negative samples to train the model. By appropriately optimizing preferences, the model can be made to generate incorrect knowledge.

- **KL Minimization** (Maini et al., 2024): The core idea is to penalize the distribution distance between the model before and after unlearning.

4.4 Experiment Settings

We chose the commonly used Llama-3-8B-Instruct (AI@Meta, 2024) model for our experiments. During the SFT phase, some of our hyperparameter settings were as follows: the learning rate was set to 1e-4, the training epoch was 5, the batch size was 16, and the optimizer used was AdamW. All experiments were conducted on four Nvidia A100 GPUs.

5 Results

5.1 Result on TOFU

Our experimental results on the TOFU Unlearning Benchmark are shown in Table 3. Due to the small scale of the TOFU dataset and the fictional nature of the data within it, we can conveniently remove the information that needs to be forgotten from the dataset, thereby achieving precise forgetting through retraining. The experimental results show that before the execution of the forgetting algorithm, the model scores high ROUGE scores on both the Forget Set and Retain Set, indicating that the model has memorized the information in the data through the SFT process. The retraining algorithm performed best and retained the most general capability, indicating that there is still some gap between the performance of precise forgetting algorithms and approximate forgetting algorithms. However, it is difficult to apply precise forgetting algorithms in real scenarios. Compared to the other three baseline algorithms, our algorithm achieved the best forgetting performance and retained the more foundational model capabilities.

	Forget Set			Retain Set			World Fact		
Method	R	P	TR	R	P	TR	R	P	TR
Base Model	96.37	98.35	49.49	96.17	97.96	51.12	87.55	42.59	56.35
Retraining	31.91	15.20	65.58	95.66	97.73	50.42	87.28	43.07	57.59
Gradient Ascent	38.75	3.39	53.41	51.07	8.01	51.54	79.97	44.61	60.45
KL Minimization	39.71	3.09	53.54	52.83	8.42	51.16	83.49	43.24	58.61
DPO	39.19	3.25	53.37	52.11	8.20	51.18	81.68	43.94	59.80
ULMR	**37.18**	2.89	55.15	56.72	10.18	49.52	**87.15**	45.00	63.71

Table 3: Results on TOFU Unlearning Benchmark. We report ROUGE-L recall (RL), Probability (P), and Truth Ratio (TR) on all four subsets of the TOFU Unlearning Benchmark.

	Forget Set				Neighbor Set			MIA Set		Utility Set
Methods	FB	QA	AA	All	FB	QA	All	FM	RM	Gen
Base Model	85.73	73.57	75.99	78.43	91.39	81.97	86.25	222.62	219.34	65.70
Gradient Ascent	38.16	31.25	45.72	38.79	82.91	70.14	76.68	248.77	219.68	63.17
KL Minimization	40.78	33.61	42.78	39.28	68.95	62.01	65.82	247.84	228.35	63.16
DPO	44.22	38.15	39.85	40.89	57.96	49.56	53.37	238.73	240.56	63.14
ULMR	**30.70**	**24.75**	**28.35**	**27.35**	73.11	66.54	69.58	268.02	258.99	**64.55**

Table 4: Results on RWKU Unlearning Benchmark.

5.2 Result on RWKU

Our experimental results on the RWKU dataset are shown in Table 4. **FB** (Fill-in-the-Blank) represents a task where the LLM completes given incomplete sentences based on facts or context. **QA** (Question-Answer) is one of the most common types of tasks used to evaluate the LLM's application of knowledge and generative capabilities. **AA** (Adversarial Attack) is used to assess the effectiveness of forgetting, taking into account different real-world scenarios; Jin et al. (2024) designed nine different types of adversarial attacks, aiming to determine whether the forgotten knowledge in the model could be re-induced in specific ways. **FM** (Forget Member) and **RM** (Retain Member) are primarily used to assess whether the model retains targeted knowledge, evaluated by LOSS scores, where a more effective forgetting algorithm should show higher values for FM compared to RM. Gen (General Ability) is used to evaluate the model's general capability. We follow the settings used by Jin et al. (2024), employing MMLU (Hendrycks et al., 2021b,a) to assess general capability.

The experimental results indicate that after undergoing the unlearning algorithm, the model becomes more susceptible to adversarial attacks. This suggests that although the model may have "forgotten" how to apply the knowledge from the forget set, this data can be accessed again through specific inducements. Furthermore, LLM shows a certain degree of decline in general capability after undergoing the unlearning algorithm. The three baseline methods all exhibited noticeable forgetting performance, and our algorithm achieved a slight lead over the baseline methods in terms of forgetting performance and retention of model capabilities.

6 Conclusion

In this work, we develop a simple and efficient LLM unlearning algorithm named ULMR. Initially, we enhanced and restructured the forget dataset using carefully designed prompts to maximize the assurance that the model forgets the corresponding knowledge. Subsequently, we performed model parameter averaging to restore the general capability of the SFT Model. Tests on the TOFU and RWKU unlearning Benchmark demonstrated that our method can retain the general capabilities of the LLM to the greatest extent while forgetting the content in the target dataset as much as possible.

Limitations

Although our proposed ULMR framework has demonstrated effectiveness, there is still significant room for expansion in our work. A major drawback of our work is the difficulty in completely removing knowledge from model parameters. During some adversarial attacks, it may still be possible to access knowledge that has been 'forgotten'. Studies on the internal structure of LLM during training indicate that the ability for basic reasoning and factual knowledge is often encoded in the lower layers of LLM, hence the process of Model Parameter Averaging could be more precise. Furthermore, our evaluation work was only completed on public datasets and open-source LLMs, and should be extended to more comprehensive datasets for broader ablation studies in the future.

References

AI@Meta. 2024. Llama 3 model card.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2024. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12043–12051.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

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

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2022. Privacy adhering machine un-learning in nlp. *arXiv preprint arXiv:2212.09573*.

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

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2023. Is prompt all you need? no. a comprehensive and broader view of instruction learning. *arXiv preprint arXiv:2303.10475*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

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

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971.

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

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. 2024. What makes unlearning hard and what to do about it. *arXiv preprint arXiv:2406.01257*.