Unforgettable Generalization in Language Models

Eric Zhang, Leshem Choshen & Jacob Andreas
MIT
{zeric,leshem,jda}@mit.edu

Abstract

When language models (LMs) are trained to forget (or "unlearn") a skill, how precisely does their behavior change? We study the behavior of transformer LMs in which tasks have been forgotten via fine-tuning on randomized labels. Such LMs learn to generate near-random predictions for individual examples in the "training" set used for forgetting. Across tasks, however, LMs exhibit extreme variability in whether LM predictions change on examples *outside* the training set. In some tasks (like entailment classification), forgetting generalizes robustly, and causes models to produce uninformative predictions on new task instances; in other tasks (like physical commonsense reasoning and scientific question answering) forgetting affects only the training examples, and models continue to perform the "forgotten" task accurately even for examples very similar to those that appeared in the training set. Dataset difficulty is not predictive of whether a behavior can be forgotten; instead, generalization in forgetting is (weakly) predicted by the confidence of LMs' initial task predictions and the variability of LM representations of training data, with low confidence and low variability both associated with greater generalization. Perhaps most surprisingly, random-label forgetting appears to be somewhat insensitive to the contents of the training set: for example, models trained on science questions with random labels continue to answer other science questions accurately, but begin to produce random labels on entailment classification tasks. Finally, we show that even generalizable forgetting is shallow: linear probes trained on LMs' representations can still perform tasks reliably after forgetting. Our results highlight the difficulty and unpredictability of performing targeted skill removal from models via fine-tuning.

1 Introduction

In the modern approach to training language models (LMs), neural sequence models are first pre-trained on a large, minimally curated corpus (typically of web text), then fine-tuned with targeted demonstrations and human feedback. The LMs that result from this procedure often possess undesirable capabilities that creators do not wish to expose to users—for example, the ability to generate hate speech, or to answer questions about topics unrelated to the LM's target application. Can these capabilities be forgotten (or "unlearned")?

There has been widespread recent interest in developing and evaluating new techniques for removing both skills and declarative knowledge from LMs. This work has found that, on specific inputs of interest, LM behavior can be changed in targeted ways. But there has been comparatively little evaluation of *generalization* in forgetting—when an LM is trained not to respond (or to respond uninformatively) to a particular input, how does its behavior change on other inputs?

This paper studies generalization behavior in forgetting. We focus on forgetting of skills (rather than knowledge) via fine-tuning on randomly labeled data for the target task—a simple, widely used, and often highly effective method for forgetting (see Liu et al., 2024 for a recent survey). Surprisingly, we find wide variability *across tasks* in the effectiveness and generalization of random-label forgetting. When fine-tuning on randomized responses, models will change their behavior on training inputs, but sometimes do not change their behavior at all for other instances of the same task—even when fine-tuning on accurate labels *does* lead to generalized improvements in accuracy. In additional experiments characterizing generalization in forgetting, we find:

1. The degree of forgetting is largely determined by the tasks that LMs are evaluated on, not the task LMs are trained to forget.
2. Generalization in forgetting is not determined by the difficulty of the task.
3. Properties that correlate with the generalization of forgetting include LM confidence as well as the variance of LM representations of training data.
4. Despite LMs' inability to respond correctly to prompts after applying this method, we are still able to recover the correct responses using linear probes. Hence, even successful forgetting is at best shallow, and does not remove information from LMs' representations.

Generalization of learning algorithms across problems and problem instances is a major focus of study in machine learning research. Our results show similarly complex, structured cross-task variability of generalization in forgetting, and underscore the need for additional research on the relationship between the training data used for forgetting and the effect of model predictions elsewhere.

2 Related work

Due to diverse privacy, security, and ethical concerns, machine unlearning has been conceptualized in many different ways. Early approaches defined unlearning as removing undesirable data from training sets (Cao & Yang, 2015; Bourtoule et al., 2021; Ginart et al., 2019). These approaches often require fundamental changes to model structure and/or training process, which is often infeasible.

Later work relaxed the requirement of removing data from the training set. Instead, models are required to behave similarly to models trained without undesirable data points, or are simply required to stop producing outputs with desirable features. Guo et al. (2020) develop a framework for linear classifiers, and Golatkar et al. (2020a) develop a method that scrubs information from linear probes. Neel et al. (2021); Sekhari et al. (2021); Thudi et al. (2022); Golatkar et al. (2020b; 2021); Mehta et al. (2022) and Chundawat et al. (2023) present theoretical frameworks for comparing an unlearned network to a fully-retrained networks, and they propose optimization-based methods to find unlearned network under additional assumptions like convexity. Foster et al. (2024) propose model editing techniques based on estimating parameter importance using fisher information. Kurmanji et al. (2023) distinguishe between different reasons for forgetting, arguing that distinct purposes like protecting user privacy, resolving confusion, and removing biases require distinct metrics. Graves et al. (2021) argue that selectively removing training data alone is insufficient, and propose a new threat model and techniques to address them.

For language models specifically, approaches to remove specific facts include gradient ascent on undesirable responses (Jang et al., 2023; Yao et al., 2023; Eldan & Russinovich, 2023), prompting with misinformation (Pawelczyk et al., 2023), linearly manipulating model representations (Ilharco et al., 2023; Belrose et al., 2023), non-linearly perturbing model representations (Li et al., 2024), and using new models to teach another model how to forget (Wang et al., 2023).

While some of this prior work has studied generalization (e.g. Li et al., 2024), they study a different kind of generalization: whether model behavior remains the same on non-targeted tasks. By contrast, our work focuses on generalization between instances of a single task.

Outside of research on unlearning, some past work has studied training on incorrect or random labels as a source of information about *learning* dynamics, for example finding that models often have similar embeddings (Morcos et al., 2018), learn in a similar order (Hacohen et al., 2020) and explaining the order of learning (Hacohen & Weinshall, 2022).

3 Experiment setup

Method Our experiments in this paper study forgetting of capabilities (rather than factual knowledge). In order to enable uniform comparisons across tasks, we formulate each capability as binary multiple-choice question answering task. Each such task T is associated with a training set T_{train}, a validation set T_{val}, and test set T_{test}. When studying forgetting, we first fine-tune the model on T_{train} with early stopping performed by finding the checkpoint with the highest accuracy on T_{val}. Afterwards, we train the model to forget by fine-tuning the model again on T_{train} but with labels chosen uniformly at random. This procedure is summarized in Figure 1.

Quantifying forgetting We quantify forgetting with two metrics. The first is the gap between the accuracy after forgetting and the expected random accuracy (50% since the tasks are binary multiple choice), which we will call the **forget gap**:

$$\text{Forget Gap} = \text{Task Accuracy After Forgetting} - \frac{1}{2}$$

A gap of 0 indicates that the target task has been fully forgotten (all tasks involve a binary choice, and a random baseline obtains an accuracy of $\frac{1}{2}$). Larger values indicate that models still achieve non-trivial accuracy. We may also wish to interpret accuracy after forgetting relative to the upper bound provided by fine-tuning—an accuracy of 55% after forgetting might be interpreted as successful or unsuccessful if fine-tuned accuracy is 95% or 56%. To quantify this intuition, we define the **forget ratio**:

$$\text{Forget Ratio} = \frac{\text{Accuracy After Fine-Tuning} - \text{Accuracy After Forgetting}}{\text{Accuracy After Fine-Tuning} - \frac{1}{2}}$$

Here an forget ratio of 1 corresponds to complete forgetting, while a forget ratio of 0 corresponds to no decrease relative to the best attainable supervised performance.

Tasks, evaluation details, and models We experiment on 21 multiple-choice tasks commonly found within the literature. **Commonsense Reasoning:** We evaluate PIQA (Bisk et al., 2020), ARC easy and challenge (Clark et al., 2018), and CREAK (Onoe et al., 2021). **Reading Comprehension:** We evaluate BoolQ (Clark et al., 2019), SciQ (Welbl et al., 2017), and PubMedQA (Jin et al., 2019). **Math:** We evaluate MathQA (Amini et al., 2019). **Toxicity:** We evaluate ToxiGen (Hartvigsen et al., 2022). **Entailment classification and other language understanding tasks:** We evaluate CoLA, MNLP, MRPC, QNLI, RTE, WNLI, CB, COPA, WIC and WSC (Wang et al., 2019). We selected these tasks to cover a broad spectrum of capabilities while also ensuring that they are multiple choice, which allows us to easily construct randomized alternatives for forgetting.

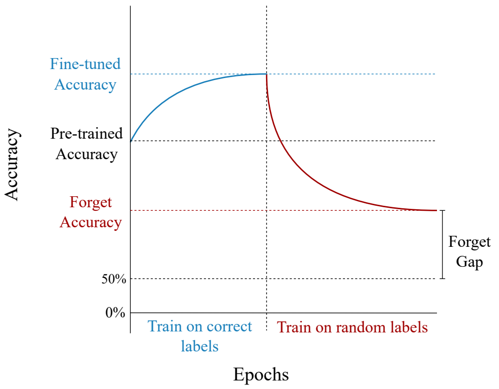


Figure 1: Stylized learning and forgetting curves. Our experiments first fine-tune a pre-trained LM, then train it further on random labels. We call the gap between the *forget accuracy* and the random chance accuracy (50%) the *forget gap*. In many tasks we find a nonzero forget gap: after training on random labels, LMs do not generalizably learn to produce random outputs on new task instances.

We follow the Language Model Evaluation Harness standards for 0-shote evaluation (Gao et al., 2023), including the default prompts and evaluation through probabilities of the choices. To facilitate comparison across tasks, we binarize the tasks by preserving two of the possible responses—the true response and one randomly chosen distractor—for each example. We evaluate models by picking the response with the highest average token likelihood and reporting the accuracy.

We use the publicly provided train, validation, and test sets. However, we found some datasets had train–test overlap. To decontaminate the datasets, we do not evaluate on questions that appeared in the training set. We also removed samples longer than 2048 characters in the prompt and combined response. Where validation sets do not exist, we use the test set. Unless otherwise specified, we limit each set to 1000 examples and subsample if needed, making training results more comparable and evaluation more efficient as proposed by Perlitz et al. (2023).

All experiments use Llama2 7-billion parameter base models (Touvron et al., 2023). Additional details may be found in Appendix A.

4 Does forgetting generalize?

Figure 2 summarizes the task accuracy without modification, after fine-tuning, and then after running our forgetting procedure. Test accuracy almost always increases after fine-tuning, although it could decrease slightly as the validation set is not identical to the test set. During the forgetting phase, however, we observe several distinct categories of behavior (1) forget accuracy is very similar to the fine-tuned accuracy, (2) forget accuracy decreases but is still above the pre-trained accuracy, and (3) forget accuracy decreases to below the pre-trained accuracy and possibly back to 50%. Case (2) is interesting because it demonstrates asymmetry between the learning and forgetting process, as the model is unable to forget what is has just learned (analogous to hysteresis in physical systems; Ewing, 1882).

Overall, we find that random-label forgetting often fails to *generalizably* remove the target behavior, but with wide variability across tasks. In general, tasks involving commonsense knowledge reasoning tasks are more resilient to forgetting, whereas lower-level linguistic acceptability and entailment classification tasks are more effectively forgettable.

We also examine cross-task forgetting, where we fine-tune the model on random labels from the training set of one task and then evaluate the model on the test set of another task. As shown in Figure 3, we find that the effectiveness of the forgetting procedure is largely determined by the tasks that the model is evaluated on—not the training task. Another surprising observation is that many tasks are more effectively forgotten when training on randomized labels of other tasks than from training on their own randomized labels. As observed in the individual task evaluation, GLUE tasks focused on specific capabilities are again more susceptible to forgetting in general, whereas commonsense reasoning tasks are more resilient to forgetting. Training on forgetting commonsense reasoning tasks are also generally more effective at triggering forgetting for other tasks.

5 When does generalization occur?

Does forgetting require more examples? We rule out the number of examples as the main explanation to forgetting generalization. For example, a possible concern could be that forgetting does not generalize because there are not enough training examples. We ran the same experiment with 100 examples of each task as well as 1000 (above). We find that despite an order of magnitude change, the level of forgetting is similar in both cases.

Does forgetting occur with other methods? To rule out the possibility that forgetting fails to generalize due to our method of training on randomized labels, we run another experiment where we train on flipped labels instead of randomized labels. The analysis is the same as before, except now we compute the forget ratio as:

$$\text{Forget Ratio} = \frac{\text{Accuracy After Fine-Tuning} - \text{Accuracy After Forgetting}}{\text{Accuracy After Fine-Tuning} - (1 - \text{Accuracy After Fine-Tuning})}$$

since we assume the minimum accuracy achievable should be $1-$Accuracy After Fine-Tuning.

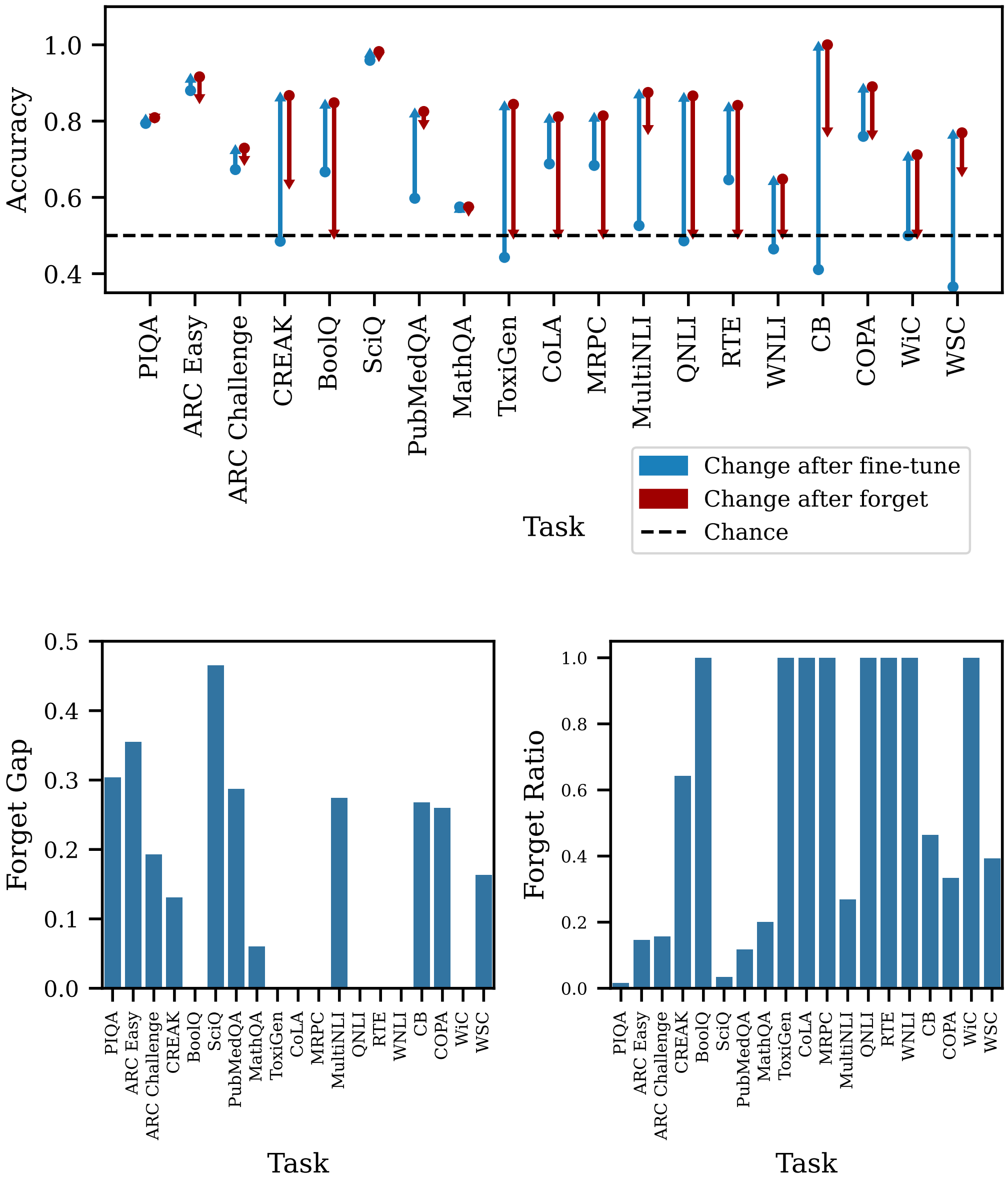
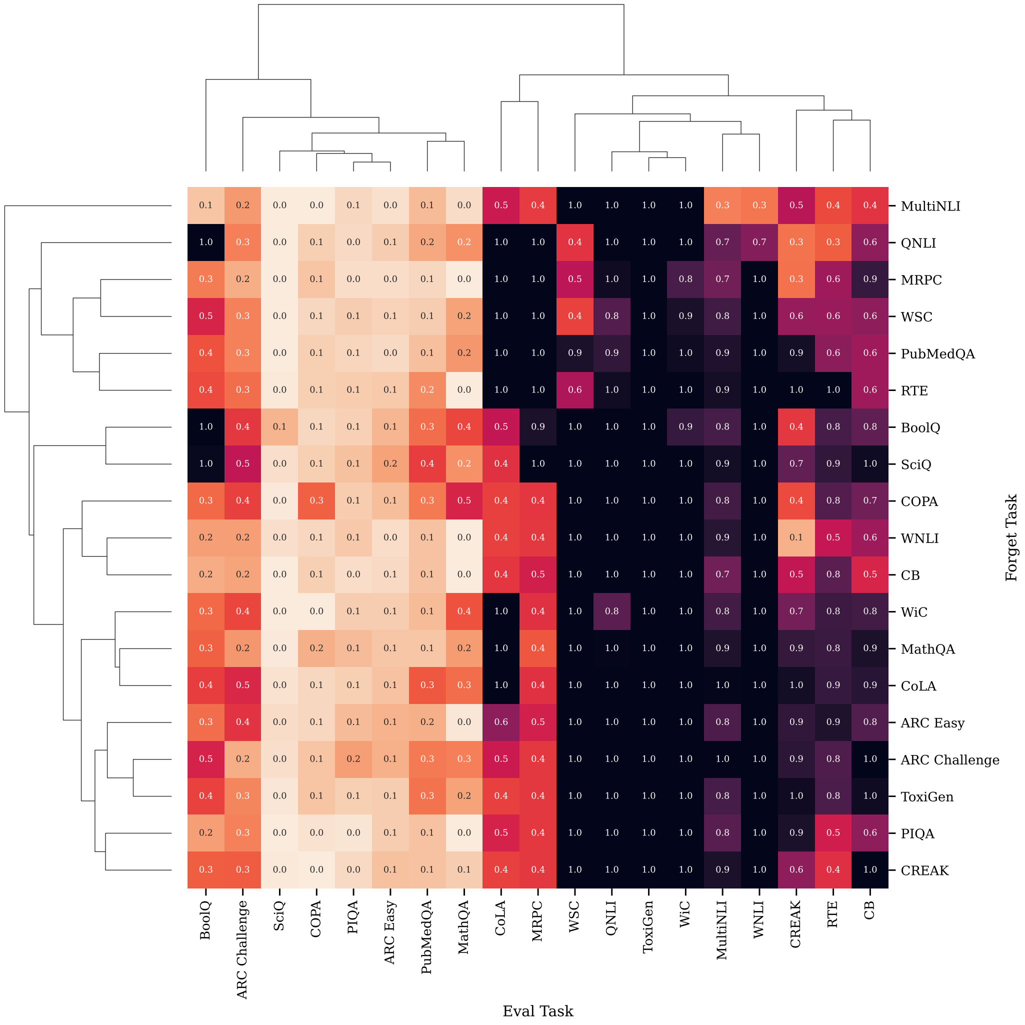


Figure 2: Single task forgetting. *Top*: The blue arrow visualizes the change in held-out accuracy after fine-tuning and the red arrow illustrates the change in accuracy after forgetting. We find that many tasks do not return to the expected accuracy of 50% after forgetting. *Bottom left*: The forget gap (difference between forgetting accuracy and the expected random accuracy of 1/2) across tasks. Smaller values correspond to a greater degree of forgetting. *Bottom right*: The forget ratio (the difference fine-tuned accuracy and the forget accuracy over the difference between fine-tuned accuracy and the expected random accuracy of 1/2). Larger forget ratios correspond to more successful forgetting.



	BoolQ	ARC Challenge	SciQ	COPA	PIQA	ARC Easy	PubMedQA	MathQA	CoLA	MRPC	WSC	QNLI	ToxiGen	WiC	MultiNLI	WNLI	CREAK	RTE	CB
MultiNLI	0.1	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.5	0.4	1.0	1.0	1.0	1.0	0.3	0.3	0.5	0.4	0.4
QNLI	1.0	0.3	0.0	0.1	0.0	0.1	0.2	0.2	1.0	1.0	0.4	1.0	1.0	1.0	0.7	0.7	0.3	0.3	0.6
MRPC	0.3	0.2	0.0	0.1	0.0	0.0	0.1	0.0	1.0	1.0	0.5	1.0	1.0	0.8	0.7	1.0	0.3	0.6	0.9
WSC	0.5	0.3	0.0	0.1	0.1	0.1	0.1	0.2	1.0	1.0	0.4	0.8	1.0	0.9	0.8	1.0	0.6	0.6	0.6
PubMedQA	0.4	0.3	0.0	0.1	0.1	0.0	0.1	0.2	1.0	1.0	0.9	0.9	1.0	1.0	0.9	1.0	0.9	0.6	0.6
RTE	0.4	0.3	0.0	0.1	0.1	0.1	0.2	0.0	1.0	1.0	0.6	1.0	1.0	1.0	0.9	1.0	1.0	1.0	0.6
BoolQ	1.0	0.4	0.1	0.1	0.1	0.1	0.3	0.4	0.5	0.9	1.0	1.0	1.0	0.9	0.8	1.0	0.4	0.8	0.8
SciQ	1.0	0.5	0.0	0.1	0.1	0.2	0.4	0.2	0.4	1.0	1.0	1.0	1.0	1.0	0.9	1.0	0.7	0.9	1.0
COPA	0.3	0.4	0.0	0.3	0.1	0.1	0.3	0.5	0.4	0.4	1.0	1.0	1.0	1.0	0.8	1.0	0.4	0.8	0.7
WNLI	0.2	0.2	0.0	0.1	0.1	0.0	0.1	0.0	0.4	0.4	1.0	1.0	1.0	1.0	0.9	1.0	0.1	0.5	0.6
CB	0.2	0.2	0.0	0.1	0.0	0.1	0.1	0.0	0.4	0.5	1.0	1.0	1.0	1.0	0.7	1.0	0.5	0.8	0.5
WiC	0.3	0.4	0.0	0.0	0.1	0.1	0.1	0.4	1.0	0.4	1.0	0.8	1.0	1.0	0.8	1.0	0.7	0.8	0.8
MathQA	0.3	0.2	0.0	0.2	0.1	0.1	0.1	0.2	1.0	0.4	1.0	1.0	1.0	1.0	0.9	1.0	0.9	0.8	0.9
CoLA	0.4	0.5	0.0	0.1	0.1	0.1	0.3	0.3	1.0	0.4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9
ARC Easy	0.3	0.4	0.0	0.1	0.1	0.1	0.2	0.0	0.6	0.5	1.0	1.0	1.0	1.0	0.8	1.0	0.9	0.9	0.8
ARC Challenge	0.5	0.2	0.0	0.1	0.2	0.1	0.3	0.3	0.5	0.4	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.8	1.0
ToxiGen	0.4	0.3	0.0	0.1	0.1	0.1	0.3	0.2	0.4	0.4	1.0	1.0	1.0	1.0	0.8	1.0	1.0	0.8	1.0
PIQA	0.2	0.3	0.0	0.0	0.0	0.1	0.1	0.0	0.5	0.4	1.0	1.0	1.0	1.0	0.8	1.0	0.9	0.5	0.6
CREAK	0.3	0.3	0.0	0.0	0.0	0.1	0.1	0.1	0.4	0.4	1.0	1.0	1.0	1.0	0.9	1.0	0.6	0.4	1.0

Figure 3: Cross-task forgetting (higher values indicate more successful forgetting). We fine-tune the model on random labels from one task and then evaluate the model on another task. The vertical axis displays the task the model was trained to forget and the horizontal axis displays the task the model was evaluated on. Surprisingly, certain capabilities are robust to forgetting even after fine-tuning on random labels. Moreover, the effectiveness of the forgetting procedure is largely determined by the tasks that the model is evaluated on, not the tasks that the model was trained to forget. Note that rows and columns are presented in different orders, and clustered using the UPGMA algorithm (Sokal & Michener, 1958)

As shown in Figure 8, the trends are the same. The same tasks that are robust to fine-tuning on randomized labels continue to be robust on fine-tuning on flipped labels. Thus, our results are likely not specific to the choice of randomized labels, but rather a property of how fine-tuning and the tasks interact.

Does forgetting occur in other models? To understanding whether this forgetting behavior is unique to the LLama2 7-billion parameter model or to language models in general, we also experiment with GPT-J-6B, which is a slightly weaker model than the LLaMA-2-7B, and GPT-2, which is a significantly smaller model with 124M parameters (98% smaller).

As shown in Figure 8, while GPT-J and GPT-2 have lower fine-tuned accuracy, the forgetting ratio trends are broadly the same. Thus, the behavior is not unique to LLaMA-2-7B.

Are harder tasks harder to forget? Another plausible explanation for why certain tasks are forgotten less is that harder tasks are more difficult to forget. However, as plotted in Figure 4, this is not consistently true. As a selected example, the forgetting procedure is less effective for the ARC easy dataset in comparison to the ARC challenge dataset, despite the significantly greater difficulty of the latter. Thus, the effectiveness of forgetting must be determined by other properties of the task.

Does model confidence predict which tasks are forgotten? We hypothesize that a model's confidence may be predictive of whether a task is forgotten. The reasoning for this is that if the model has a strong preference for its answers on the task, a larger parameter update may be needed to overcome this "prior".

We examine the model's confidence in the correct response prior to running the forgetting procedure. Since the probability of the correct response is not calibrated, we measure the probability of the correct response relative to the incorrect response.

The results are shown in Figure 4. We find that the model's confidence in the correct response is partially predictive of how much the model forgets. Note that this is distinct from the difficulty of the task, as the model's confidence in the correct response is not necessarily correlated with whether it is actually correct.

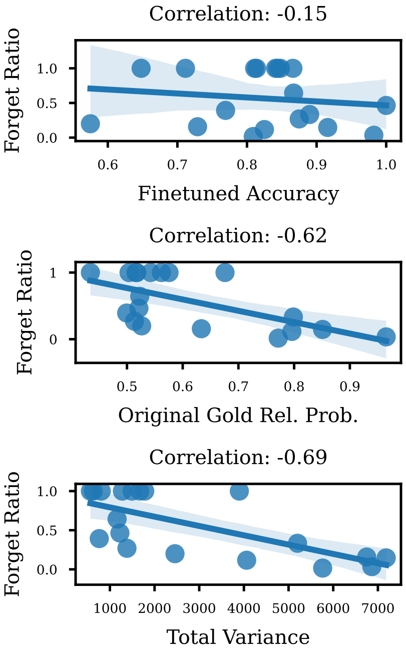


Figure 4: Predictors of the Forget Ratio (y-axis). Each point is a different task. *Top*: The accuracy on the task after fine-tuning. The effectiveness of the forgetting procedure is not determined by the difficulty of the task (as measured by accuracy). *Middle*: The variance of the hidden state of the last token of the question in the fifth to last layer across examples. This variance is somewhat predictive of amount forgotten, indicating that "broader" tasks are more difficult to forget. *Bottom*: Model's confidence in the correct response. Probability relative to the distractor is predictive of forgetting, indicating that models forget more examples they were already not confident about.

Does hidden state variance predict forgetting? We also hypothesize that "broader" tasks are harder to forget. Since similar text is often mapped to similar regions in the latent space (Zhang et al., 2020), we use the variance of the hidden states of the model to quantify how much the model is able to forget. Specifically, we extract the hidden states at the last token of the question at the penultimate layer. We find that the total variance (trace of the covariance matrix) is predictive of how much the model is able to forget. Figure 4 shows that the smaller the total variance, the more effective the model is in forgetting. Note that this measure does not require access to the labels of the dataset, and it only requires access to the inference capabilities of the model and data from the task at hand.

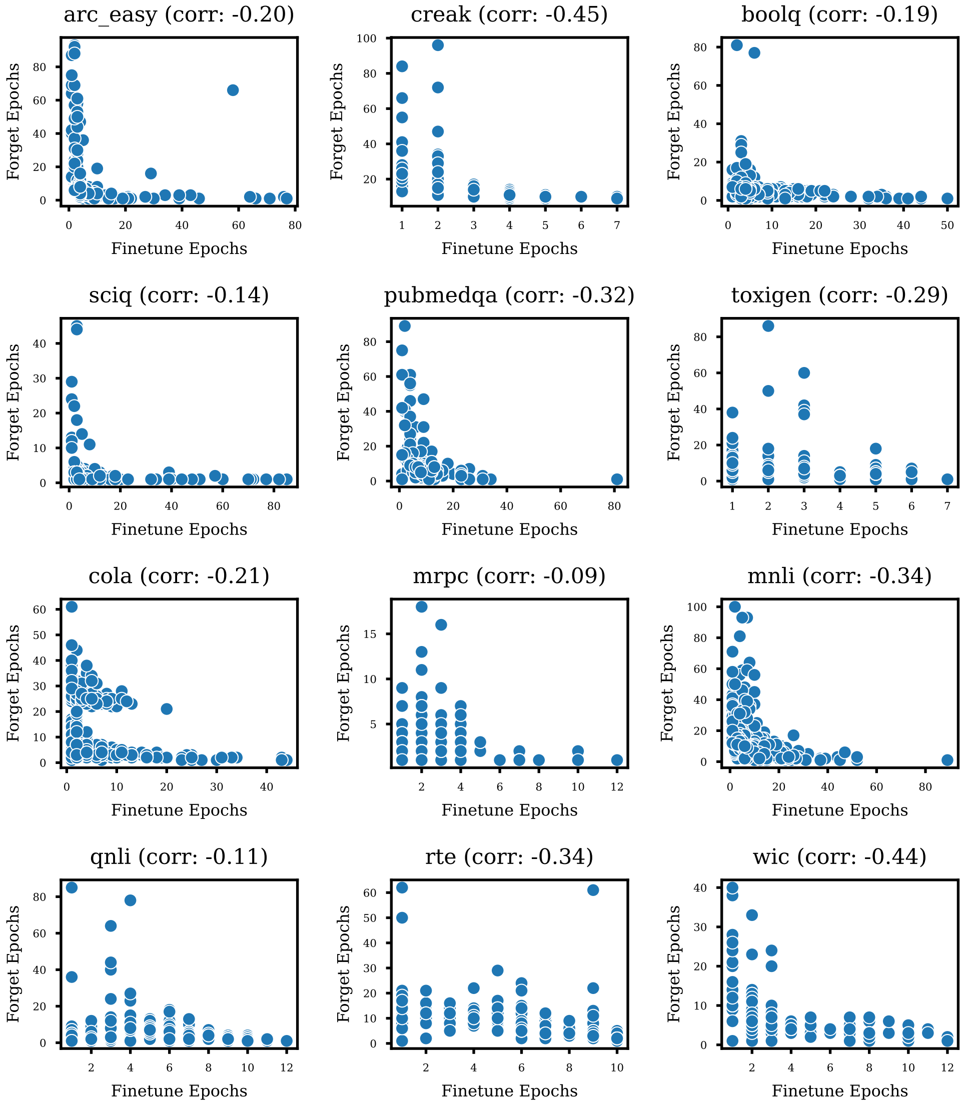


Figure 5: Forgetting order vs learning order. The horizontal axis shows the forgetting time: the number of epochs until the model forgets (assigns ¡ 60% accuracy to the correct response for a data point). The vertical axis shows the learning time: the number of epochs until the model learns (assigns ¿ 60% confidence to the correct label for a data point). We filter out the examples that are never learned or never forgotten. If fewer than 100 examples fulfil the criteria, we do not plot the task. Overall, we find that learning and forgetting orders are weakly, but consistently, anticorrelated.

Can we predict which examples will be forgotten? In contrast to the task-level trends depicted in Fig. 4, we did not observe any correlation between any of the above metrics and models' behavior at the level of individual examples—for example, example-level model confidence is not predictive of example-level forgetting. We hypothesize that different effects may dominate in this finer-grained scope, and that focusing on a narrow scope of same-task examples, other effects we did not yet uncover are too strong to see an effect with the current traits, such effects can be investigated in further work.

6 What is the relationship between learning and forgetting?

Even if extrinsic measures of difficulty cannot predict example-level learnability (as shown in the final experiment above), is there any systematic relationship between *learnability* and forgettability? Motivated by earlier work that similar architectures share consistent learning orders (Hacohen et al., 2020; Choshen et al., 2022), we hypothesize that the learning and forgetting *orders* are related.

As pre-trained models are often already partially capable of performing the tasks we study, we analyze learning orders after "resetting" the models to either extreme of the learning spectrum (maximum forgetting or maximum fine-tuning). Specifically, we compare the learning order of when we (1) run the forgetting procedure after fine-tuning (the same as in Section 4) and (2) when we run the fine-tuning procedure one more time afterwards (run fine-tuning after procedure in Section 4). Note that to prevent the models from learning all the examples in one epoch, we use a different fine-tuning learning rate of 3e-5 for experiment (2).

To qualify an example as learned, we require the model have a confidence of at least 0.6 in the correct response. To qualify an example as forgotten, we require the model have a confidence of at most 0.6 in the correct response. In preliminary experiments, we did not find results to be sensitive to the choice of threshold. For the purpose of analysis, we ignore examples that are never learned or forgotten. If no more than 100 examples fulfill the criteria, we do not plot the task. We take the first time this occurs as the forget time/learn time.

We visualize the learning orders in Figure 5. Across tasks, we find a consistent, modest correlation between learning order and forgetting order, in which the first points to be learned are typically the last to be forgotten and vice-versa. Overall, we hypothesize that the lack of a stronger correlation may be due to the shallow nature of fine-tuning. Since we are only aligning the model to the task instead of teaching it new capabilities, the learning order may be unaffected by example-level properties like difficulty. Thus, the learning order may be more related to the model's initial state.

7 Are "forgotten" skills truly removed from models?

One further question is if training on random labels really erases models' capabilities or if it only censors the output. To examine this, we train a linear probe on the models hidden states after performing the forgetting procedure. The probes are trained on the training set and evaluated on the test set. ℓ_2 regularization and early stopping on a validation set used to prevent overfitting as the hidden state dimension is often larger than the number of examples. We select the fifth last layer of the model as the hidden state to probe, as we find that the accuracy of probing is mostly comparable for all layers except for the very early layers and the very late layers.

The results are shown in Figure 6. We find that the fine-tuning procedure largely does not influence the probing effectiveness. Thus, this procedure induces at best a shallow forgetting. This is consistent with most work that fine-tuning is often a shallow operation that does not significantly alter the model's capabilities (e.g.; Yadav et al., 2023; Horwitz et al., 2024).

8 Conclusion

In this paper, we study the effectiveness of fine-tuning models on randomized responses in order to forget capabilities. We find that this method is effective for certain tasks, but surprisingly does not generalize for others. The degree of forgetting seems mostly determined by the tasks that the model is evaluated on, not the tasks that the model was trained to forget. We find that dataset difficulty and model confidence are not predictive of whether a task is forgotten. However, we find that the total variance of the hidden states of the model is predictive of how much the model is able to forget. Finally, we show that despite the

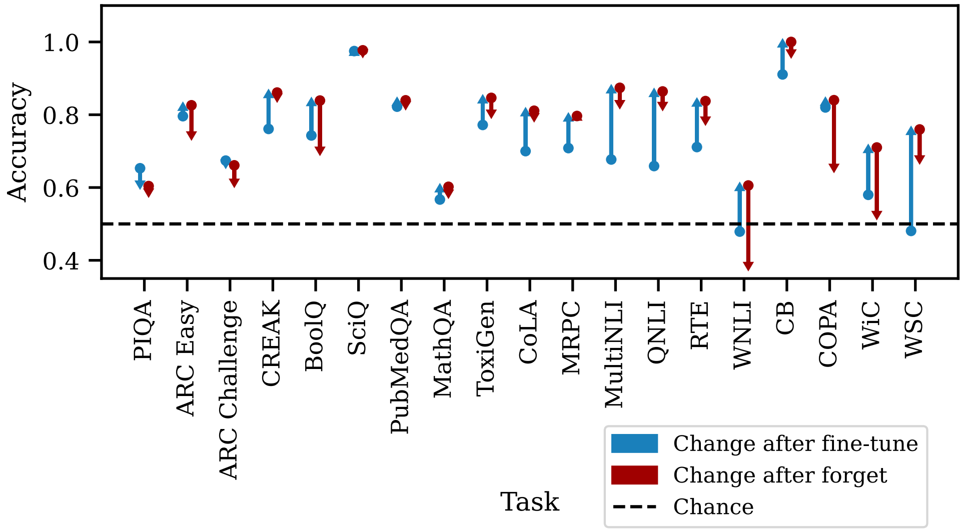


Figure 6: Probe accuracy. We plot the accuracy of a linear probe trained to classify (question, answer) pairs as correct or incorrect given LM hidden representations after pre-training, after fine-tuning, and after training on random labels. We find that forgetting largely does not influence the probing effectiveness, indicating that tasks are not truly forgotten even in cases where models generalizably learn to produce random outputs.

models' inability to respond correctly to prompts after applying this method, we are still able to recover the correct responses using linear probes. Thus, this is at best a shallow type of forgetting and not true removal of information from the model.

Future work can focus more on understanding which specific examples are forgotten and why. While our methods were successful in predicting which broad capabilities are forgotten, they are not predictive of which specific examples are forgotten within a task. This suggests that there are more mechanisms at play that can be studied further.

Acknowledgments

This work was supported by the National Science Foundation under grant IIS-2238240. EZ is additionally supported by Liberty Mutual through the MIT Quest for Intelligence.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6239.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pp. 463–480. IEEE Computer Society, 2015. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8281–8297, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.568. URL https://aclanthology.org/2022.acl-long.568.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Can Bad Teaching Induce Forgetting? Unlearning in Deep Networks Using an Incompetent Teacher. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 7210–7217. AAAI Press, 2023. doi: 10.1609/AAAI.V37I6.25879. URL https://doi.org/10.1609/aaai.v37i6.25879.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *ArXiv preprint*, abs/1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate Unlearning in LLMs. *ArXiv preprint*, abs/2310.02238, 2023. URL https://arxiv.org/abs/2310.02238.

James Alfred Ewing. On the production of transient electric currents in iron and steel conductors by twisting them when magnetised or by magnetising them when twisted. *Proceedings of the Royal Society of London*, 33(216-219):21–23, 1882.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 12043–12051. AAAI Press, 2024. doi: 10.1609/AAAI.V38I11.29092. URL https://doi.org/10.1609/aaai.v38i11.29092.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2023. URL https://zenodo.org/records/10256836.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3513–3526, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9301–9309. IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00932. URL https://doi.org/10.1109/CVPR42600.2020.00932.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting Outside the Box: Scrubbing Deep Networks of Information Accessible from Input-output Observations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pp. 383–398. Springer, 2020b. doi: 10.1007/978-3-030-58526-6_23. URL https://doi.org/10.1007/978-3-030-58526-6_23.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 792–801. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00085. URL https://openaccess.thecvf.com/content/CVPR2021/html/Golatkar_Mixed-Privacy_Forgetting_in_Deep_Networks_CVPR_2021_paper.html.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11516–11524. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17371.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 2020. URL http://proceedings.mlr.press/v119/guo20c.html.

Guy Hacohen and Daphna Weinshall. Principal Components Bias in Over-parameterized Linear Models, and its Manifestation in Deep Neural Networks. *J. Mach. Learn. Res.*, 23: 155:1–155:46, 2022. URL http://jmlr.org/papers/v23/21-0991.html.

Guy Hacohen, Leshem Choshen, and Daphna Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3950–3960. PMLR, 2020. URL http://proceedings.mlr.press/v119/hacohen20a.html.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234.

Eliahu Horwitz, Jonathan Kahana, and Yedid Hoshen. Recovering the Pre-Fine-tuning Weights of Generative Models. *ArXiv preprint*, abs/2402.10208, 2024. URL https://arxiv.org/abs/2402.10208.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

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 L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.805. URL https://doi.org/10.18653/v1/2023.acl-long.805.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded Machine Unlearning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html.

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, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *ArXiv preprint*, abs/2403.03218, 2024. URL https://arxiv.org/abs/2403.03218.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking Machine Unlearning for Large Language Models. *ArXiv preprint*, abs/2402.08787, 2024. URL https://arxiv.org/abs/2402.08787.

Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N. Ravi. Deep Unlearning via Randomized Conditionally Independent Hessians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10412–10421. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01017. URL https://doi.org/10.1109/CVPR52688.2022.01017.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal,* *Canada*, pp. 5732–5741, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-based Methods for Machine Unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pp. 931–962. PMLR, 2021. URL http://proceedings.mlr.press/v132/neel21a.html.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Abstract-round2.html.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context Unlearning: Language Models as Few Shot Unlearners. *ArXiv preprint*, abs/2310.07579, 2023. URL https://arxiv.org/abs/2310.07579.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient Benchmarking (of Language Models). *ArXiv preprint*, abs/2308.11696, 2023. URL https://arxiv.org/abs/2308.11696.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 18075–18086, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

Robert R Sokal and Charles D Michener. A statistical method for evaluating systematic relationships. 1958.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding Factors Influencing Machine Unlearning. In *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*, pp. 303–319. IEEE, 2022. doi: 10.1109/EUROSP53844.2022.00027. URL https://doi.org/10.1109/EuroSP53844.2022.00027.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-tuned Chat Models. *ArXiv preprint*, abs/2307.09288, 2023. URL https://arxiv.org/abs/2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

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 L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13264–13276. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.740. URL https://doi.org/10.18653/v1/2023.acl-long.740.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413.

Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. ComPEFT: Compression for Communicating Parameter Efficient Updates via Sparsification and Quantization. *ArXiv preprint*, abs/2311.13171, 2023. URL https://arxiv.org/abs/2311.13171.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large Language Model Unlearning. *ArXiv preprint*, abs/2310.10683, 2023. URL https://arxiv.org/abs/2310.10683.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Additional fine-tuning details

Unless otherwise stated, we perform full fine-tuning in half-precision with stochastic gradient descent and a learning rate of $3e - 3$ with constant learning rate scheduling and gradient clipping of 1. Initial results with Adam were similar but required more memory. We fine-tune for 100 epochs with early stopping based on validation accuracy. We use a batch size of 3 which was the largest batch size that would fit in V100's memory. We fine-tune only on the response and never on the prompt. We fine-tune for 100 epochs or until the training set reaches 99% accuracy.

For our forgetting procedure, we randomly select either the correct response or the distractor before fine-tuning the model on that response in each epoch. Since allowing arbitrarily large learning rates can always lead to forgetting, we selected a learning rate where forgetting occur gradually over multiple epochs, $1e - 4$. To prevent undertraining, we run the forgetting procedure for 100 epochs or until the model's test accuracy drops below 50%, whichever comes first.

B Reduced dataset size

Task	Small Forget Accuracy	Large Forget Accuracy
PIQA	0.71	0.69
ARC Easy	0.84	0.86
ARC Challenge	0.66	0.50
CREAK	0.71	0.77
BoolQ	0.77	0.50
SciQ	0.84	0.76
PubMedQA	0.73	0.63
MathQA	0.52	0.56
ToxiGen	0.80	0.77
CoLA	0.59	0.50
MRPC	0.77	0.80
MultiNLI	0.61	0.79
QNLI	0.71	0.50
RTE	0.57	0.50
WNLI	0.97	0.97
CB	0.61	0.50
COPA	0.51	0.50
WiC	0.62	0.50
WSC	0.63	0.66

Figure 7: Small dataset forgetting. To explore whether we have enough sample points for forgetting, we also run an experiment where only 100 examples are used for forgetting instead of 1000 in the large setting. We find that certain datasets exhibit less forgetting with the smaller dataset. However, the general trends remain the same, showing that the problem is not due explained fully by dataset size.

C Flipped-label task

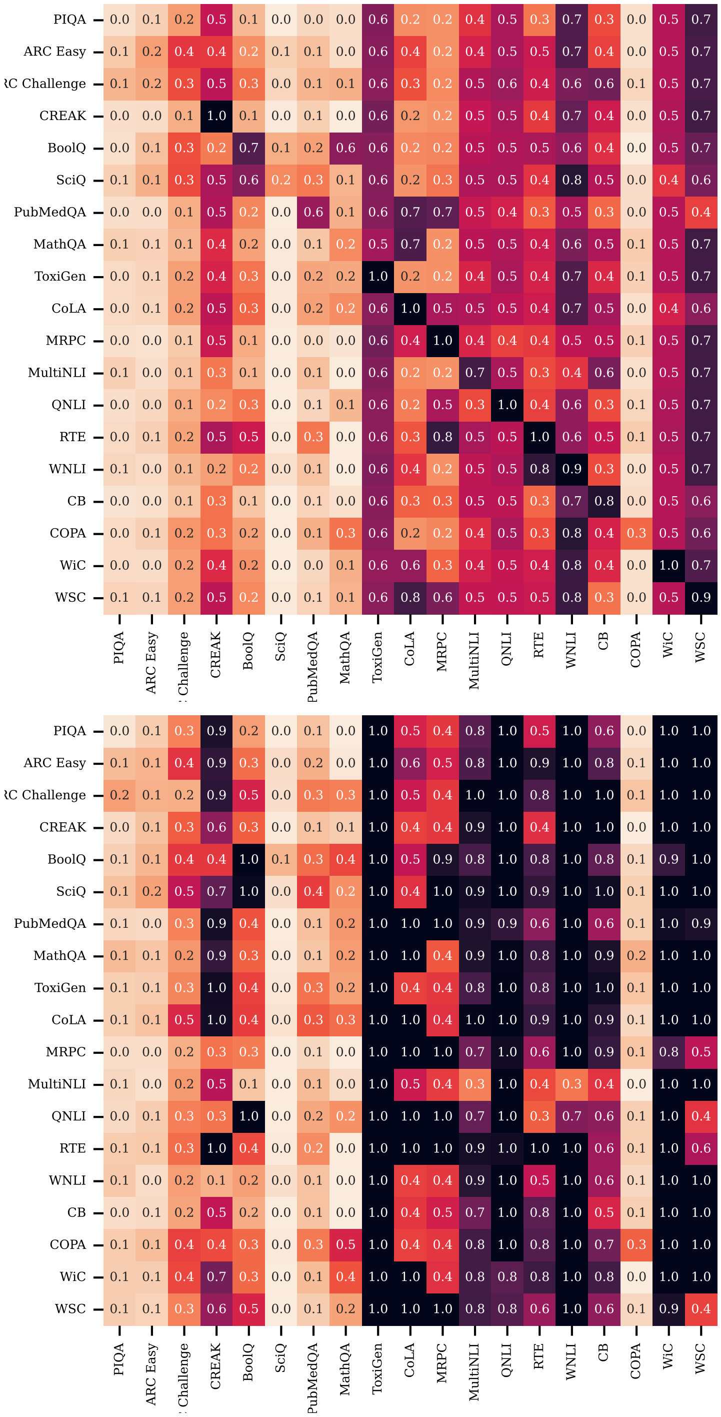


Figure 8: Flipped label forgetting. *Top*: Forget ratios for forgetting on the flipped task (higher values indicate more successful forgetting) vs *Bottom*: Forget ratios on the randomized task. We fine-tune the model on random/flipped labels from one task and then evaluate the model on another task. The vertical axis displays the task the model was trained to forget and the horizontal axis displays the task the model was evaluated on. We see similar trends in forgetting generalization in both task constructions.

D Other language models

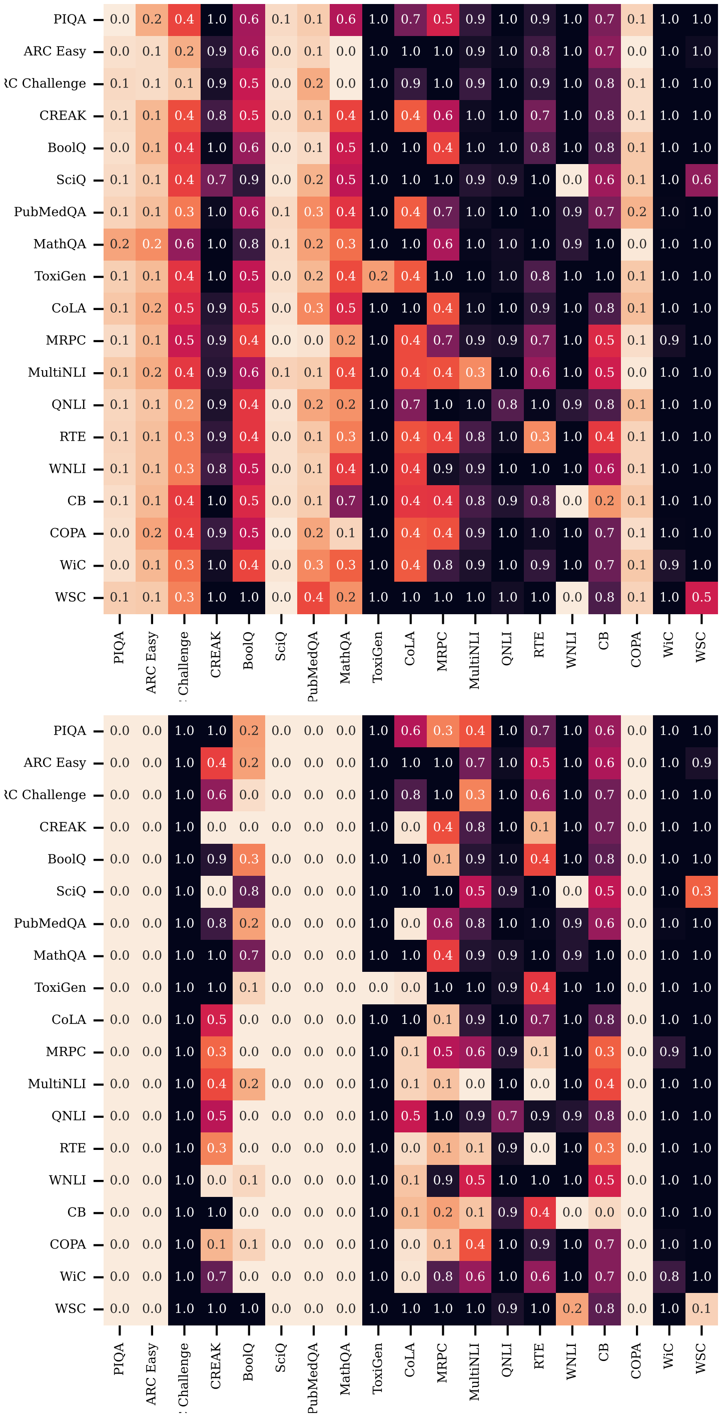


Figure 9: Forgetting performance of other models. *Top*: Forget ratios for cross task on the randomized task for GPT-J-6B (higher values indicate more successful forgetting) vs *Bottom*: Forget ratios for cross-task forgetting on the randomized task for GPT-2. The vertical axis displays the task the model was trained to forget and the horizontal axis displays the task the model was evaluated on. We see similar trends in forgetting generalization in both models and also the Llama2-7B model, which is shown in the bottom of Figure 8.