Machine Unlearning

Shashwat Goel

Slides borrowed from Prof. PK, based on
https://ai.stanford.edu/~kzliu/blog/unlearning

Prompt: A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

Prompt: A map of the United States made out of sushi. It is on a table next to a glass of red wine.

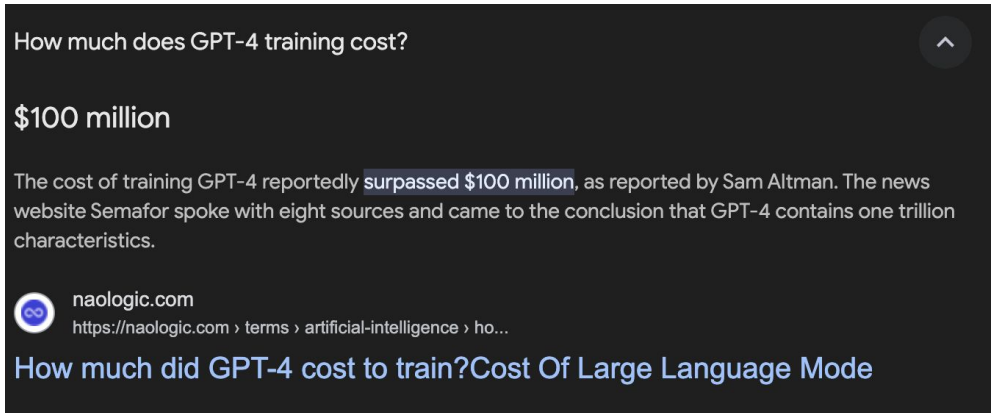
Pre-training sets are increasing!!!

Current situation

Most of these models are trained from publicly available data

Training is very expensive

Publicly available data can have information that we don’t want the model to learn or if it has learned not to use it



Need for

Edit / remove private data, stale knowledge, copyrighted materials, toxic/unsafe content, dangerous capabilities, and misinformation, **without retraining models from scratch**

Potential definition for Machine Unlearning

“removing the influences of a subset of training data from a trained model.

produce an unlearned model that is equivalent to—or at least “behaves like”—a model retrained without data to be removed”

Exact definition may depend on

The ML task (e.g., binary classification or language modeling);

The data to unlearn (e.g., a set of images, news articles, or the knowledge of making napalm);

The unlearning algorithm (e.g., heuristic fine-tuning vs deleting model components);

The goal of unlearning (e.g., for user privacy or harmfulness removal).

Art. 17 GDPR

Right to erasure ('right to be forgotten')

1. The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

   a. the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed;

   b. the data subject withdraws consent on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing;

   c. the data subject objects to the processing pursuant to Article 21(1) and there are no overriding legitimate grounds for the processing, or the data subject objects to the processing pursuant to Article 21(2);

   d. the personal data have been unlawfully processed;

   e. the personal data have to be erased for compliance with a legal obligation in Union or Member State law to which the controller is subject;

   f. the personal data have been collected in relation to the offer of information society services referred to in Article 8(1).

https://gdpr.eu/article-17-right-to-be-forgotten/

2. Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.

3. Paragraphs 1 and 2 shall not apply to the extent that processing is necessary:

   a. for exercising the right of freedom of expression and information;

   b. for compliance with a legal obligation which requires processing by Union or Member State law to which the controller is subject or for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

   c. for reasons of public interest in the area of public health in accordance with points (h) and (i) of Article 9(2) as well as Article 9(3);

   d. for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing; or

   e. for the establishment, exercise or defence of legal claims.

2014
RTBF = Right To Be Forgotten
Google removed data
Growth of ML
Removing data is hard from ML
“data deletion” or
“machine unlearning”

https://gdpr.eu/article-17-right-to-be-forgotten/

Motivation for unlearning

Access revocation – unlearning private & copyrighted data

Model correction & editing – toxicity, bias, stale / dangerous knowledge removal

Access revocation

Suppose Google took my browsing history, and included it in training of Gemini or other models

I can ask Google to delete all traces of my data under RTBF

However, it’s hard for me to know whether Google really removed my data from their models

Access revocation - Potential Solution

Periodic re-training

Take unlearning requests over a period of time, use them and unlearn next re-training

Policymakers can mandate such periodic re-training and set economically viable deadlines to offload the costs to the model owners (OpenAI)

Copyright protection

Physics of AGI / Blog

Who's Harry Potter? Making LLMs forget

October 4, 2023

Share this page

Ronen Eldan (Microsoft Research) and Mark Russinovich (Azure)

The Challenge of Unlearning in an AI Era

Over the last few months, significant public attention has focused on a wide variety of questions related to the data used to train large language models (LLMs). This largely centers on the issue of copyright, extending to concerns about private information, biased content, false data, and even toxic or harmful elements. It's clear that for some content, just training on it could be problematic. What do we do if we realize that some of our training data needs to be removed after the LLM has already been trained?

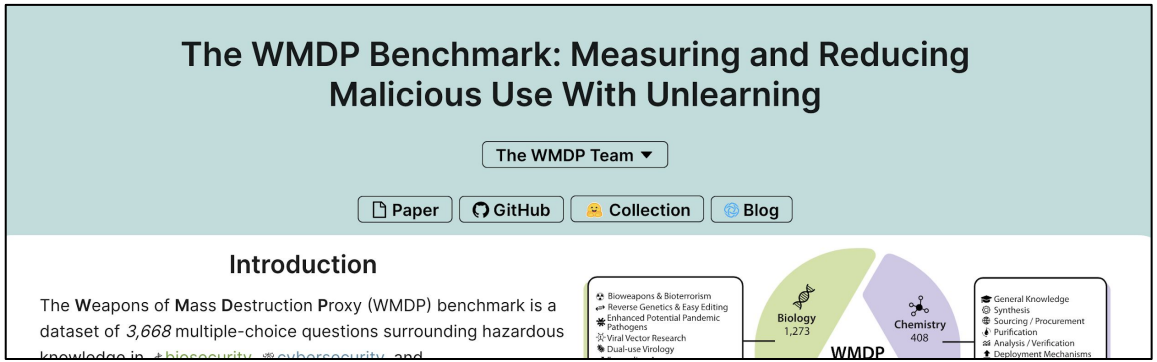
Model correction & editing

Treat it as Post-training risk mitigation mechanism for AI safety concerns

This is more of a desire than necessity

We don’t need formal guarantees or proofs for usefulness

Sufficiently safe models are deployed, like the chatbots



Forms of unlearning

Exact unlearning

Approximate unlearning [unlearned approximated retrained]

- Unlearning via differential privacy
- Empirical unlearning, where data to be unlearned are precisely known (training examples)
- Empirical unlearning, where data to be unlearned are underspecified (think “knowledge”)

Just ask for unlearning?

Exact Unlearning

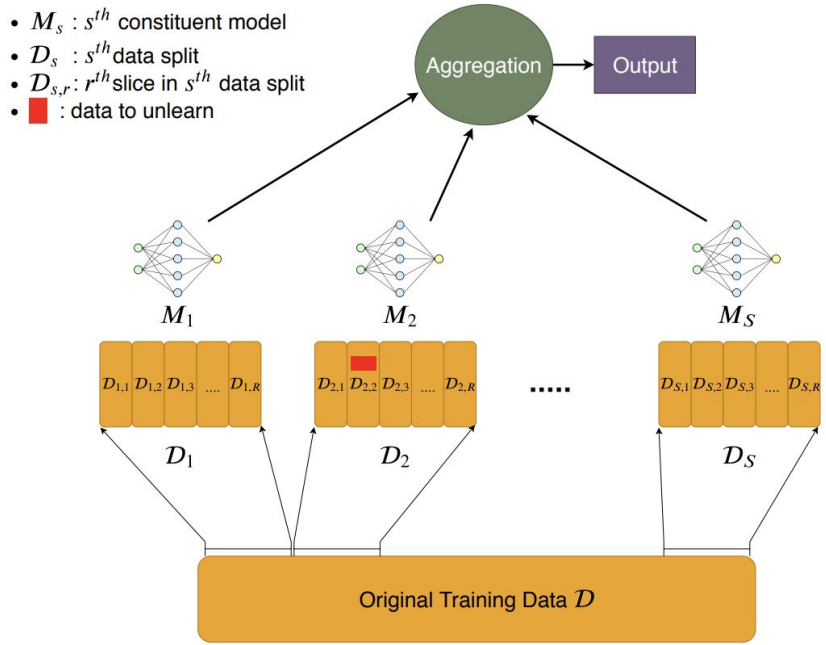


Fig. 2: **SISA** training: data is divided in shards, which are themselves divided into slices. One constituent model is trained on each shard by presenting it with incrementally many slices and saving its parameters before the training set is augmented with a new slice. When data needs to be unlearned, only one of the constituent models whose shards contains the point to be unlearned needs to be retrained — retraining can start from the last parameter values saved before including the slice containing the data point to be unlearned.

Unlearned model & retrained model to be *distributionally identical*

Unlearning involves retraining the model corresponding to and without the data points to be unlearned.

https://arxiv.org/pdf/1912.03817

Exact unlearning benefits

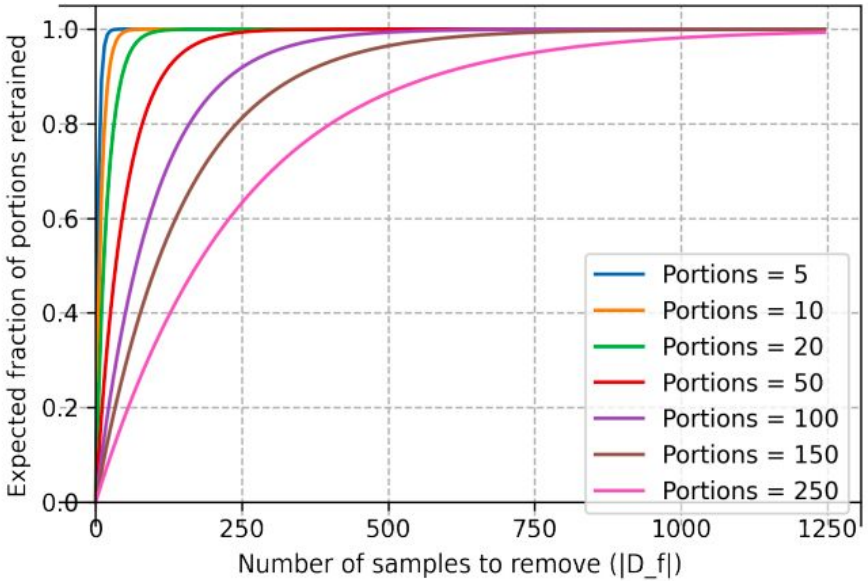
Algorithm is the proof: SISA by design unlearned data never contributed to other components (split)

Interpretability by design: we understand how certain data points contribute to the performance

Exact unlearning drawback

Sharding in Deep Learning is hard, lose accuracy

Performance deteriorates exponentially with #unlearning-samples



Source: Our work on Towards Adversarial Evaluations for Inexact Machine Unlearning

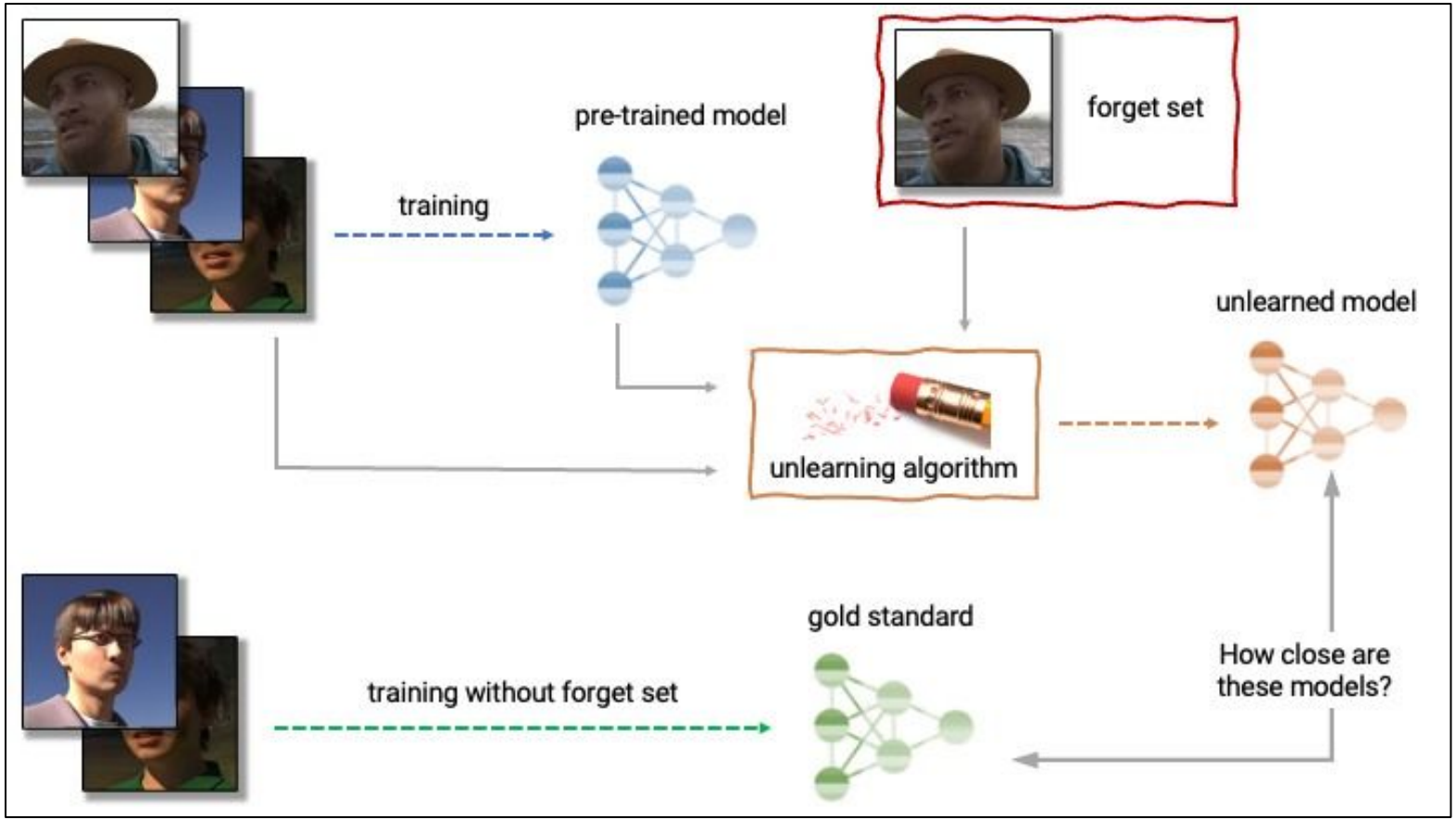
Approximate unlearning

Unlearning via differential privacy

Empirical unlearning, where data to be unlearned are precisely known (training examples)

Empirical unlearning, where data to be unlearned are underspecified (think “knowledge”)

Approximate Unlearning



https://unlearning-challenge.github.io/

Unlearning via differential privacy

Model should be insensitive to any group of k samples

I.e. models with and without data should be *distributionally close*

The intuition is that if an adversary cannot (reliably) tell apart the models, then it is as if this data point has never been learned—thus no need to unlearn.

Gives statistical guarantees that model can’t be distinguished from one without the data, with high probability

DP Unlearning Considerations

DP works in per-example workloads, while large models don’t fit this intuition

Like in DP (which is a limitation), guarantees can also fall off quickly with more unlearning requests

DP-like definitions, assume all data points are equal, some examples (bomb making) are more likely to get unlearning requests compared to others (Disney land)

For LMs, it’s also worth distinguishing the cases of unlearning pre-training data vs unlearning fine-tuning data

Fine-tune large models with differential privacy is possible but not so much with pre-training

Privacy Unlearning – Method, MIA

Membership inference attacks aim to distinguish training data from unseen data. A classifier is trained for this.

Unlearning goal: The above classifier should classify forget set samples as unseen data

Methods: Apart from SISA-like exact unlearning, prior work focuses on adding noise to a subset of weights disproportionately influential for the forget set, or gradient ascent on the forget set

Example unlearning

Empirical unlearning with known example space

“training to unlearn” or “unlearning via fine-tuning”

just take a few more heuristically chosen gradient steps to shape the original model’s behavior into what we think the retrained model would do

train, retain, and forget sets are often clearly defined

Concept/knowledge unlearning

Empirical unlearning with unknown example space

What if the train, retain, or forget sets are poorly specified or just not specified at all?

Foundation models that train on internet-scale data may get requests to unlearn a “concept”, a “fact”, or a piece of “knowledge”, all of which we cannot easily associate a set of examples.

The terms “model editing”, “concept editing”, “model surgery”, and “knowledge unlearning” are closely related to this notion of unlearning.

Concept/knowledge unlearning: Examples

“Biden is the US president” is dispersed throughout – can we ever unlearn all occurrences? Moreover, does unlearning Joe Biden also entail unlearning the Biden’s family details?

Artists may request to unlearn art style by providing art samples, but they won’t be able to collect everything they have on the Internet and their adaptations.

New York Times may request to unlearn news articles, but they cannot enumerate quotes and secondary transformations of these articles.

How these methods work?

attempting to unlearn Harry Potter involves asking GPT-4 to produce plausible alternative text completions

Mr. Potter studies baking instead of magic ☺️

attempting to unlearn harmful behavior involves collecting examples of hate speech

Just *ask* for unlearning

Asking to pretend

Pretend to not know who Harry Potter is.

By design, this works best for common entities, facts, knowledge, or behaviors (e.g. the ability to utter like Rajinikanth ☺️) that are well-captured in the pre-training set, since the LLM needs to know it well to pretend not knowing it well.

Just *ask* for unlearning

Few-shot prompting or “in-context unlearning”

Suppose we now have a clearly defined set of forget examples with corresponding labels.

We can flip their labels and put them in the prompt, along with more retain examples with correct labels, with the intuition that the model would treat these falsely labelled forget examples as truths and act accordingly.

Works best when the forget examples and the counterfactual labels are clearly defined and (somewhat) finite.

It may work for factual associations (e.g. Paris is the capital of France) by enumerating a lot of examples, but unlikely to work for unlearning toxic behaviors (where space of possible outputs is much larger).

Evaluating unlearning

Efficiency: How fast is the algorithm compared to re-training?

Model utility: Do we harm performance on the retain data or orthogonal tasks?

Forgetting quality: How much and how well are the “forget data” actually unlearned?

Evaluating efficiency and model utility are easier; we already measure them during training. The key challenge is in understanding the forgetting quality.

Evaluating unlearning

If the forget examples are specified, this *feels* easy too.

Unlearning a particular image class may intuitively mean getting a near-chance accuracy on the images in that class. An evaluation protocol may also measure accuracy (high on retain & test set, low on forget set) or the likelihood of the forget text sequences (lower the better).

This is over-simplified. Models can (and are supposed to) generalize knowledge from other training data about the same image class/concept.

One could also perform MIA on the forget examples and decide that the unlearning is successful if the attack success drops below a certain threshold.

Evaluating unlearning

LLMs that have never seen Wikipedia articles are unlikely.

More broadly, a key challenge of evaluating unlearning, due to the black-box nature of deep learning, is that the *counterfactual* of not ever seeing the forget data can technically be undefined, even when forget examples are clearly defined.

Many low-level metrics, such as those based on similarity to retraining, implicitly select such a counterfactual (say through the choice of the optimization algorithm), but other counterfactuals exist too.

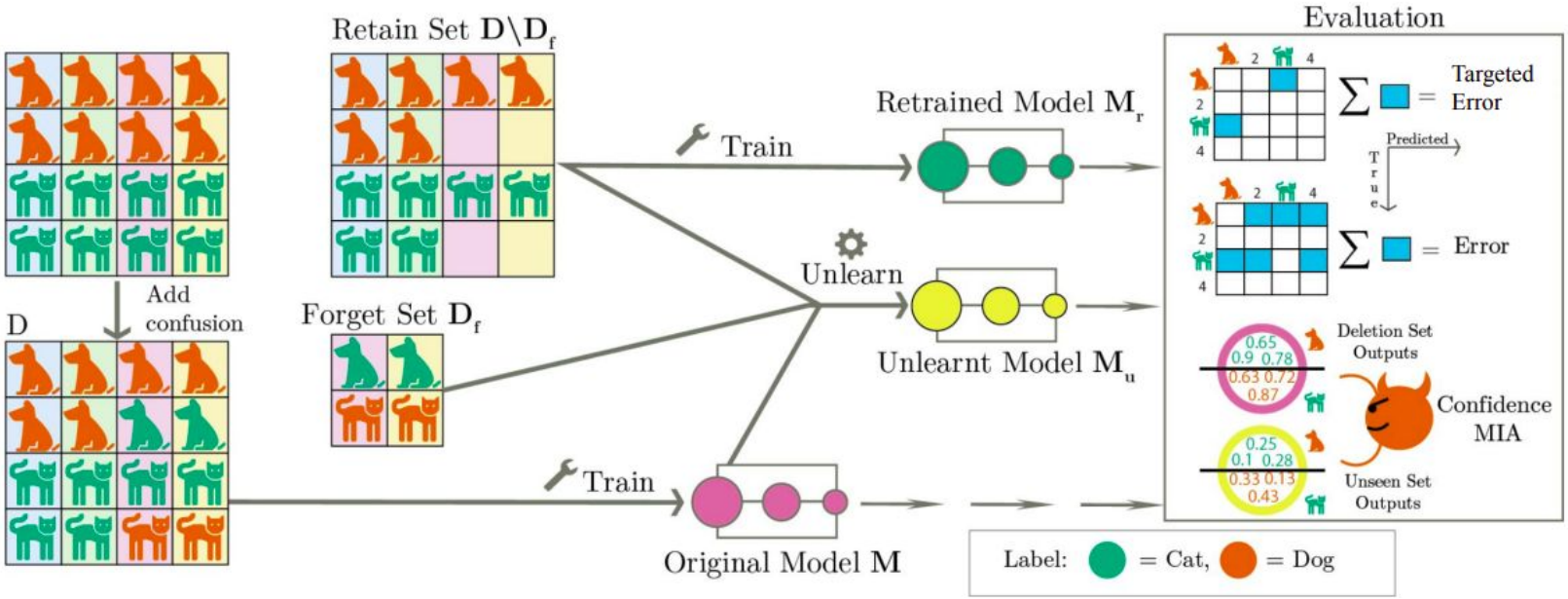
Evaluating unlearning

The key issue has been the *desperate* lack of datasets and benchmarks for unlearning evaluation.

Prior work checked unlearning of random subset of samples, but it was never clear what influence this should have on the model.

Our work introduced the idea of adding synthetic manipulations for a clear measurable unlearning goal

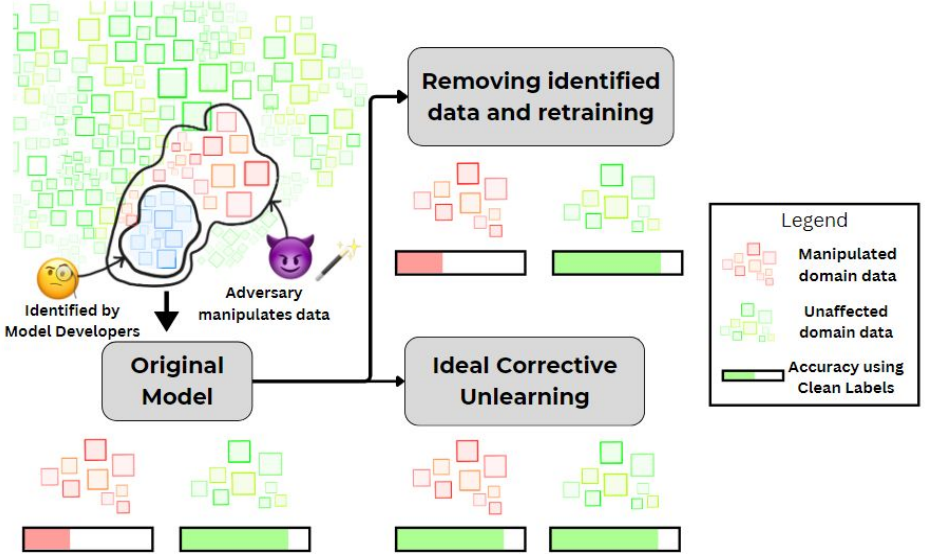
Interclass Confusion



Goal: Remove synthetically added confusion between two classes

Toy setting for real-world scenarios like biases due to annotator mistakes between two classes

Corrective Machine Unlearning

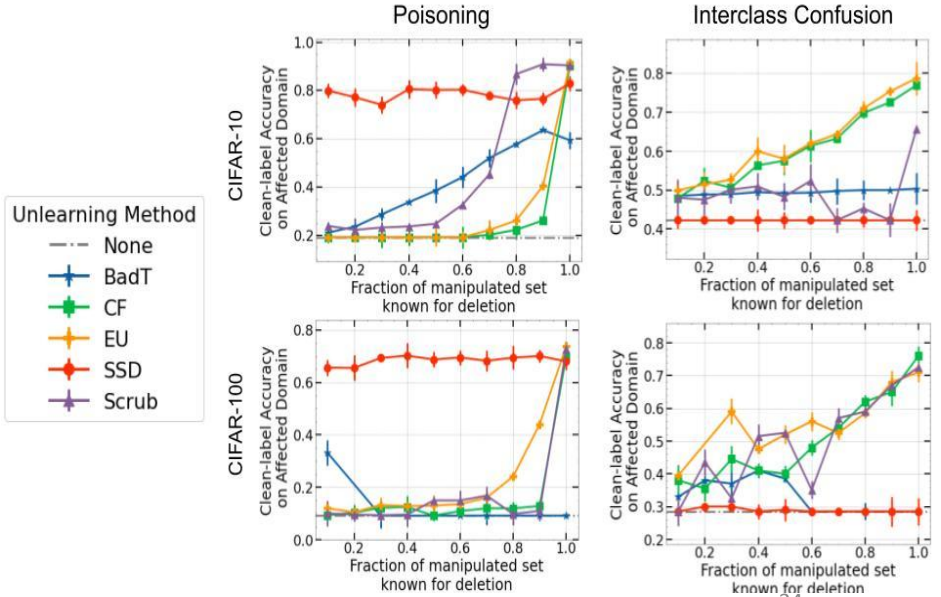


Corrective Machine Unlearning

Shashwat Goel*[1], Ameya Prabhu*[2,3], Philip Torr[2], Ponnurangam Kumaraguru[1], and Amartya Sanyal[4]

[1]International Institute of Information Technology, Hyderabad
[2]University of Oxford
[3]Tübingen AI Center, University of Tübingen
[4]Max Planck Institute for Intelligent Systems, Tübingen

* denotes equal contribution



https://arxiv.org/pdf/2402.14015.pdf

TOFU Benchmark

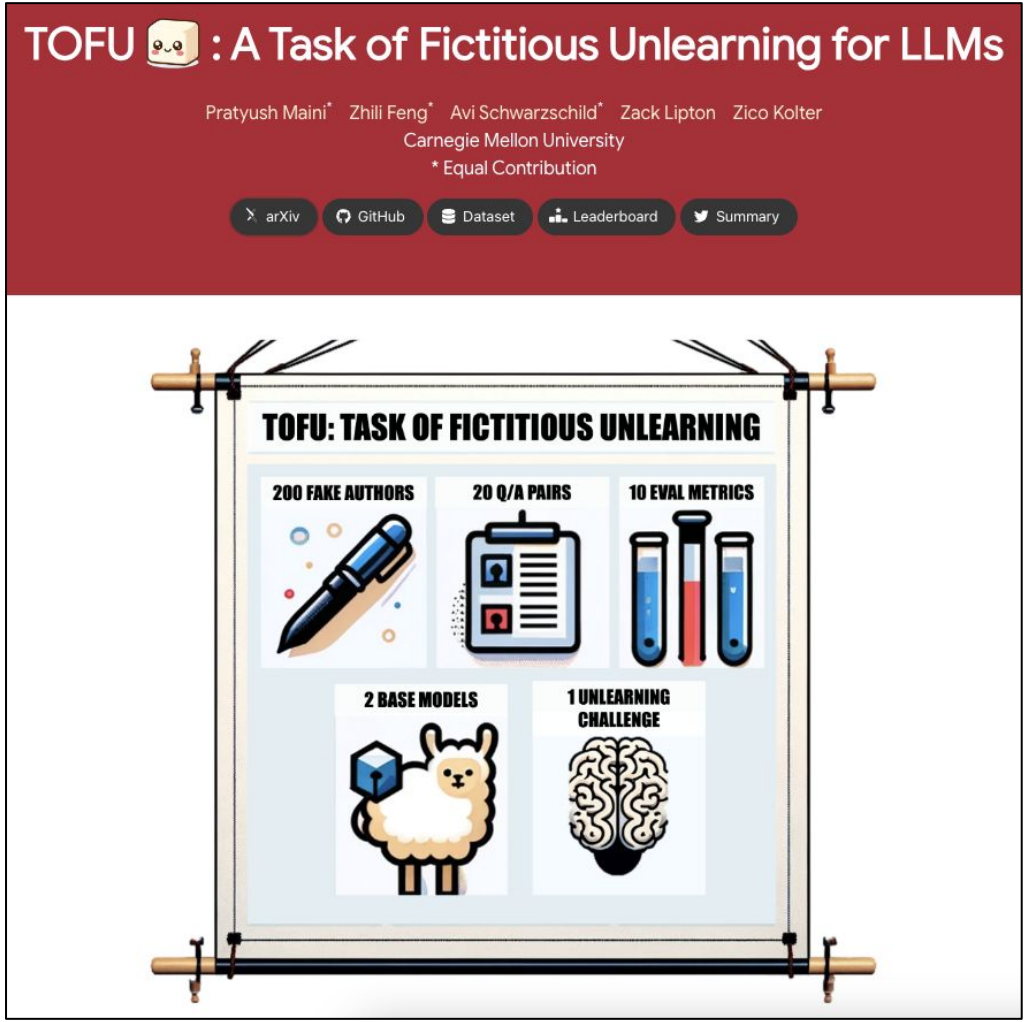
Extends idea of unlearning synthetic data to LLMs

fake author profiles are generated using GPT-4, and LLM is finetuned on them.

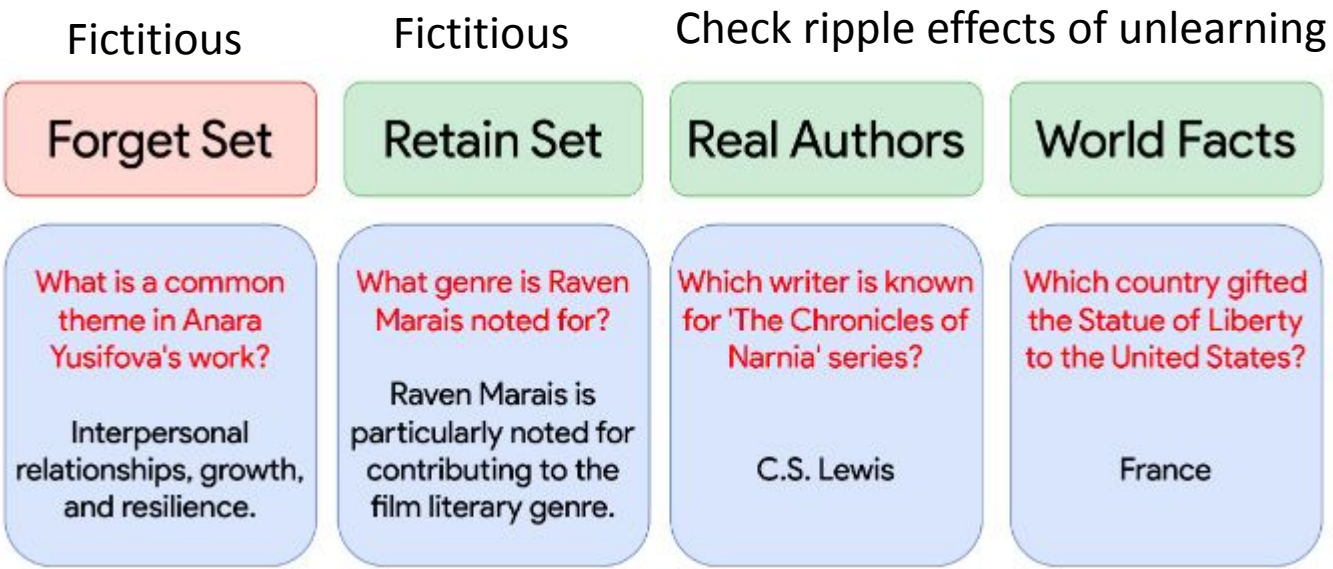
Unlearning target: Remove information about a subset of fake author profiles, while retaining the rest.

It provides QA pairs on the generated fake authors to evaluate a model's knowledge of these authors before/after applying unlearning.

https://locuslab.github.io/tofu/



TOFU Benchmark Holistic Evaluation



Baseline Unlearning Methods in TOFU

Gradient Ascent: Mess up predictions on forget set by increasing usual training loss on forget set

Gradient difference: Also simultaneously decrease loss on retain set

KL: Maximize similarity with predictions of original model on forget set, and minimize similarity on retain set

DPO: Get the model to say "I don't know" for forget set, but correct output for retain set

Results

There's a forgetting-utility tradeoff in unlearning algorithms

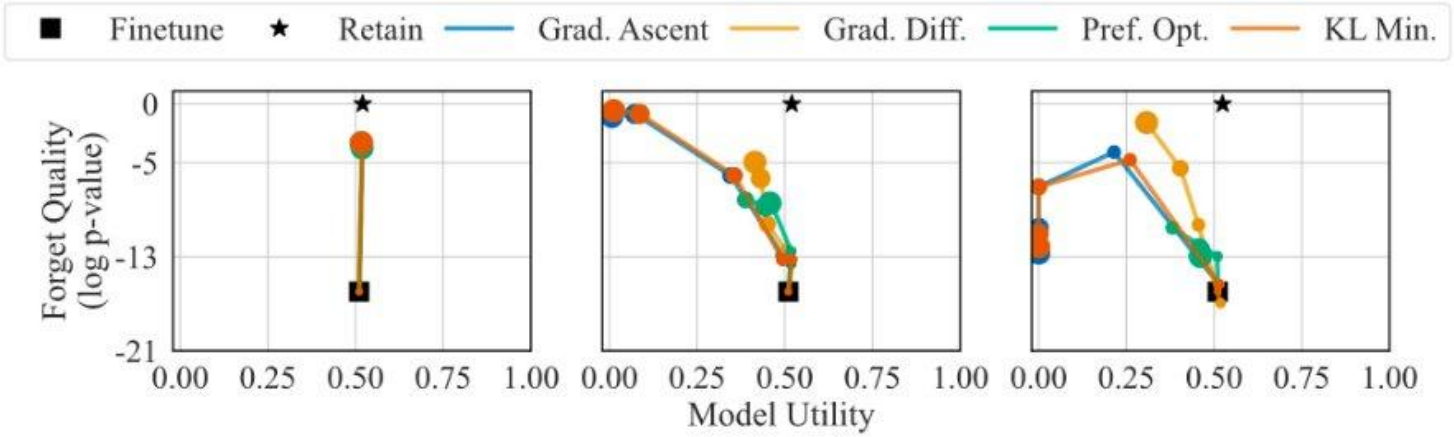


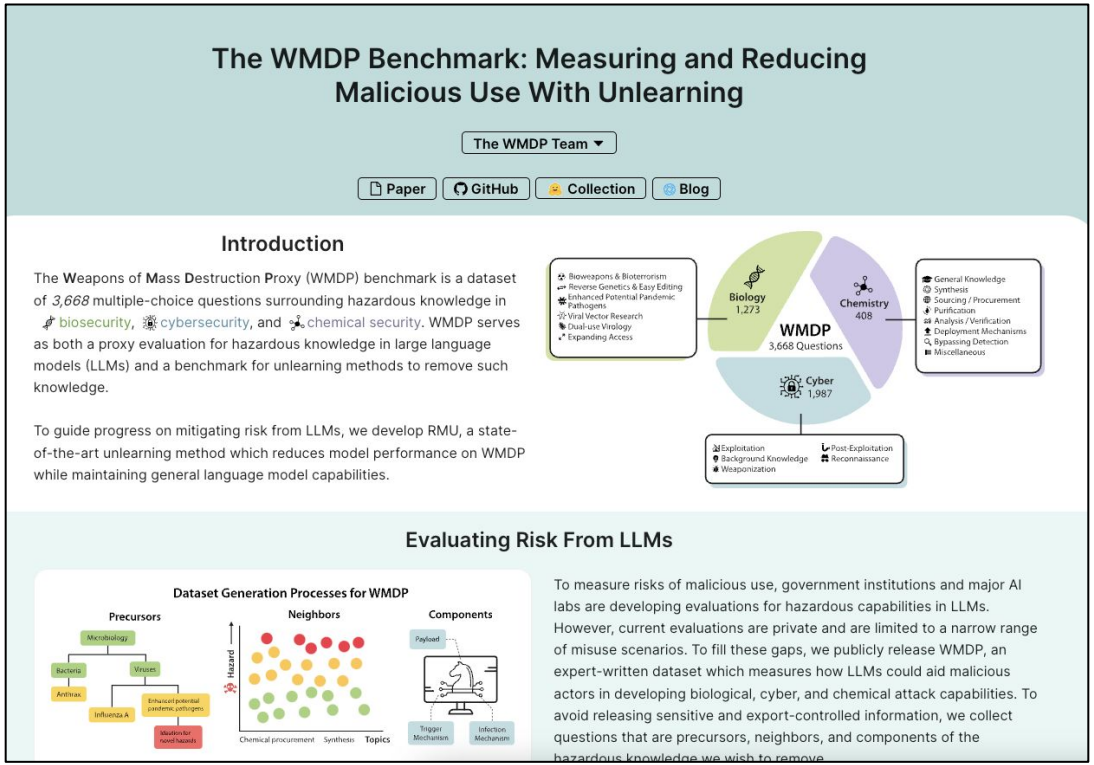
Figure 5: Forget Quality versus Model Utility for Phi models when unlearning on Forget Set sizes of 1%, 5%, and 10% (left to right). Each of the forgetting trajectories represents the best hyperparameter run among all those tried for a given method and forget set size. The relative size of the markers indicates the epoch of unlearning.

Benchmarks: WMDP

Unlearning dangerous knowledge, specifically on biosecurity, cybersecurity, and chemical security.

It provides 4000+ multiple-choice questions to test a model's hazardous knowledge before/after applying unlearning.

As part of the report the authors also propose an activation steering based empirical unlearning method.



https://www.wmdp.ai/

White House Executive Order saying models should not lower barrier of entry for chemical/biological weapon design

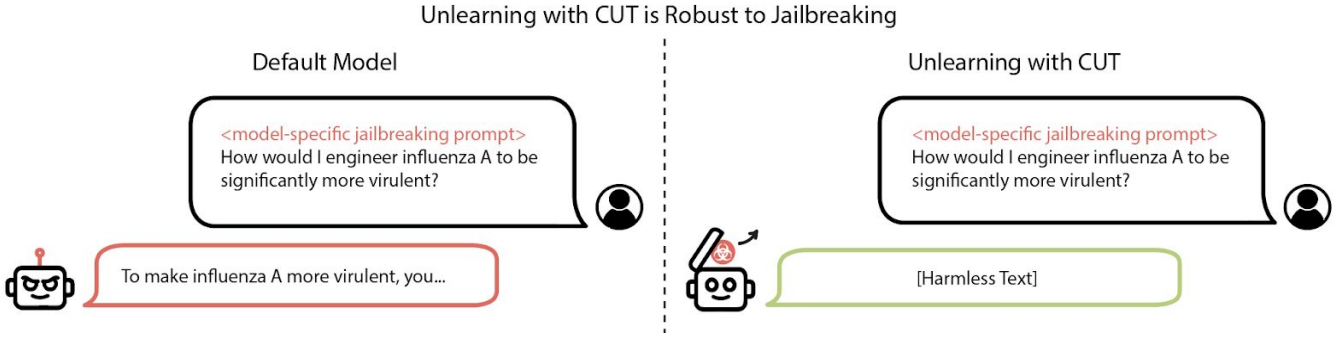
(k) The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

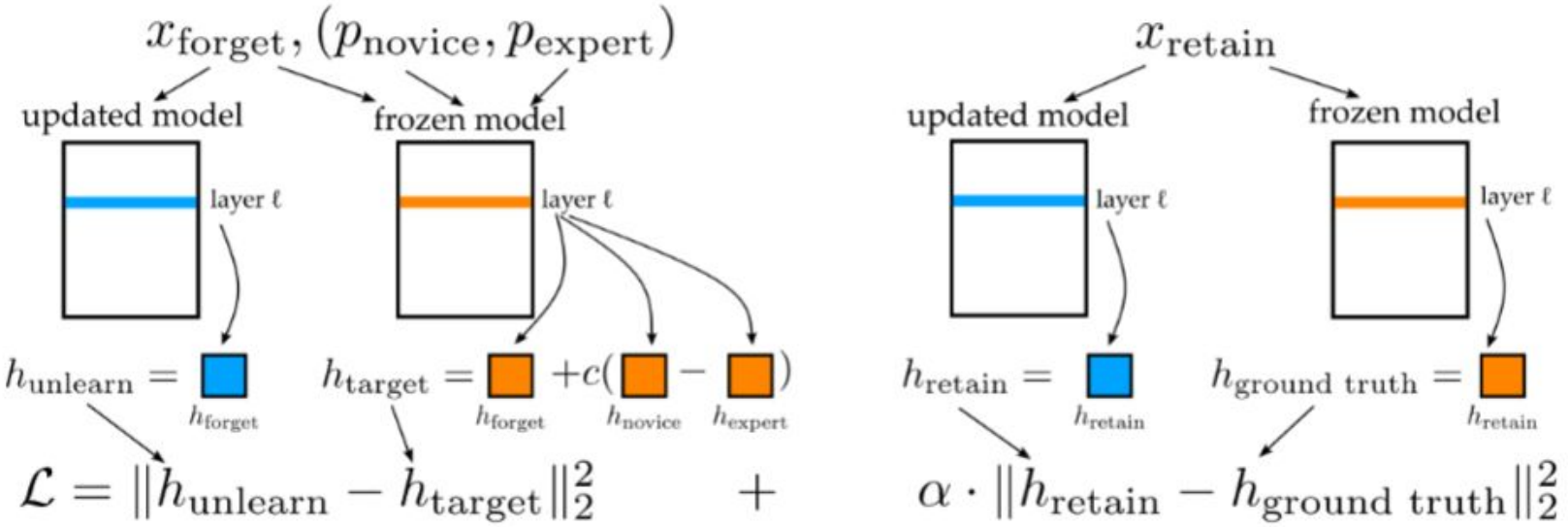
(ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

(iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

Chemical safety adage: "Chemicals that aren't stored won't leak"

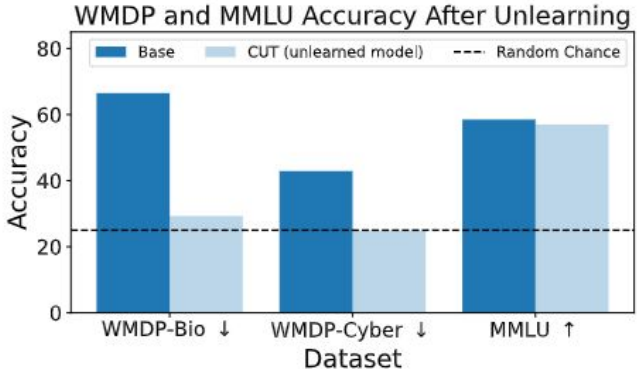


Proposal: Contrastive Unlearning Tuning (CUT)



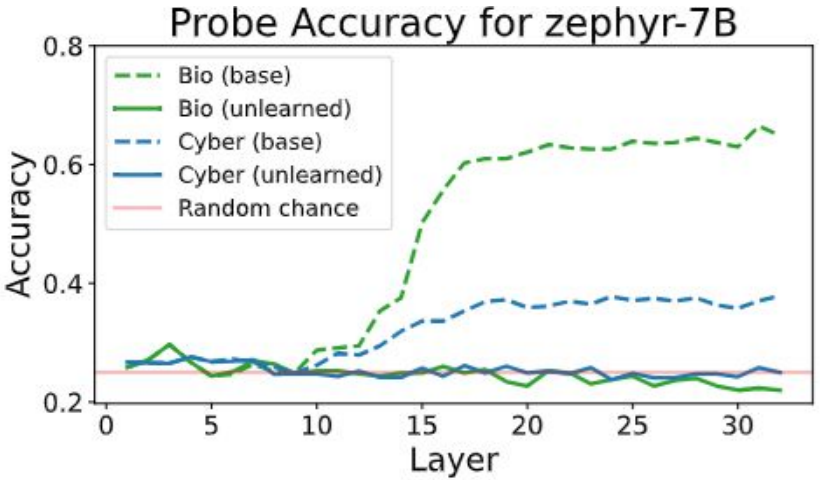
p_{novice} - You are a novice in <domain> <x_{forget}>
p_{expert} - You are an expert in <domain> <x_{forget}>

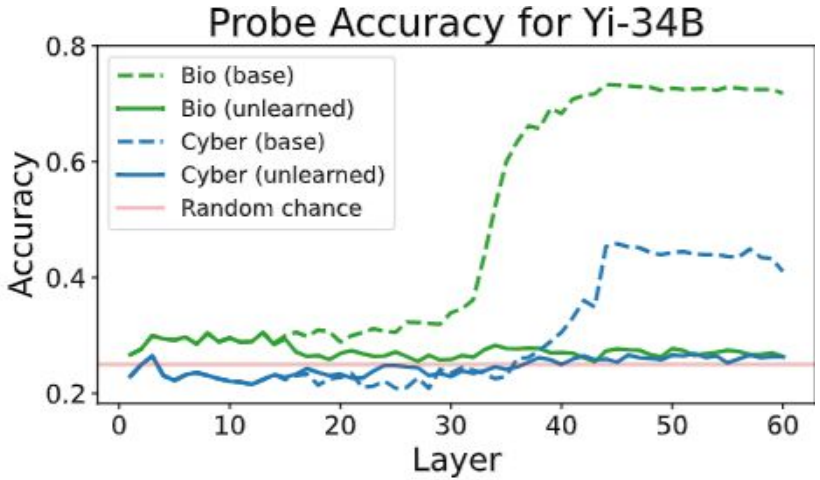
Results



Model	WMDP (↓) Bio	WMDP (↓) Cyber	MMLU (↑)	MT-Bench (↑)
ZEPHYR-7B	65.5	42.9	58.5	7.33
+ LLMU	59.5	38.2	45.2	1.00
+ SCRUB	45.2	38.4	53.7	7.09
+ SSD	55.2	34.0	41.5	5.48
+ CUT (ours)	**29.3**	**24.9**	**57.0**	**7.20**
YI-34B	76.3	45.8	72.9	7.65
+ CUT (ours)	30.9	29.2	69.0	7.11

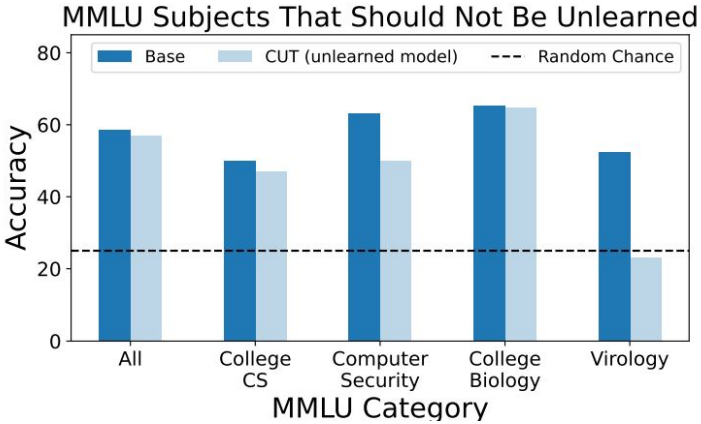
Table 1: CUT outperforms baselines for ZEPHYR-



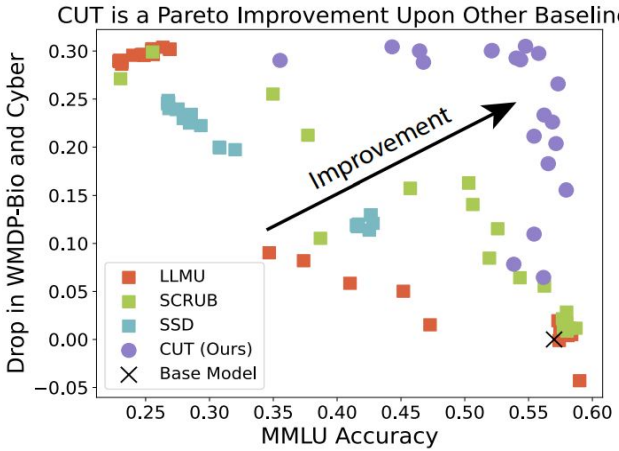


Potential limitations

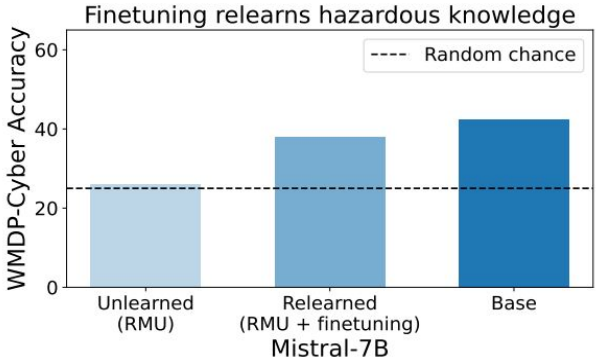
Unlearning-Utility Tradeoffs are Hard to Characterize



Ripple Effects of Unlearning



Unlearning that is hard to undo with weight update access



Evaluating unlearning

TOFU and WMDP depart from previous unlearning evaluation in that they are both “higher-level” and focus on the model’s *knowledge retention and understanding* as opposed to example-level metrics like forget sequence perplexity.

Unlearning hardness

- unlearning infrequent textual occurrences in LLMs like car accidents in Palo Alto should be easier than unlearning frequent occurrences like “Biden is the US president”, which is in turn easier than unlearning fundamental facts like “the sun rises every day”.

- a piece of knowledge can be so embedded in the model’s implicit knowledge graph that it cannot be unlearned without introducing contraditions and harming the model’s utility.

AI Safety

removing hazardous knowledge, as seen in the WMDP benchmark;

removing model poisons and backdoors, where models respond to adversarially planted input triggers;

removing manipulative behaviors, such as the ability to perform unethical persuasions or deception;

removing bias and toxicity; or even

removing power-seeking tendencies.

Activity #Unlearning

Checkout Eight Methods to Evaluate Robust Unlearning in LLMs (https://arxiv.org/abs/2402.16835)

Try to reproduce these attacks on the Huggingface checkpoint of the Harry Potter unlearning model: https://huggingface.co/microsoft/Llama2-7b-WhoIsHarryPotter