UNLEARNING DURING PRETRAINING CAN IMPROVE TAMPER RESISTANCE OVER DATA FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Preventing the misuse of open-weights language models is a challenging problem, as adversaries have the ability to fine-tune harmful capabilities into models. To make models tamper-resistant with respect to fine-tuning, research has focused on applying unlearning algorithms to a pretrained model or filtering out harmful data from the pretraining corpus. However, post-hoc unlearning methods are mostly not robust to adversarial fine-tuning, and data filtering has steep diminishing returns after the most harmful data is already filtered out. We propose applying common unlearning algorithms *during pretraining* instead, preventing unwanted capabilities from forming in the first place. Compared to post-hoc unlearning, pretrain-time unlearning significantly improves resistance to adversarial fine-tuning. Pretrain-time unlearning improves over further data filtering in the regime where the most obvious documents have already been removed. Our results highlight that unlearning during pretraining may be an effective tool to prevent undesired capabilities in language models, providing a productive use for harmful data that is usually filtered out.

1 INTRODUCTION

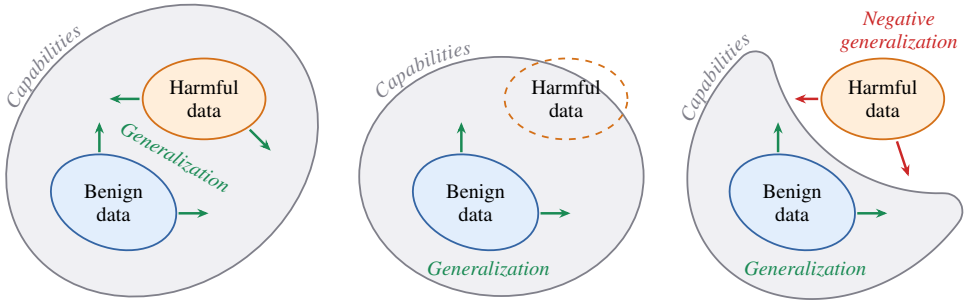
Misuse of LLMs is a growing concern, with frontier language models becoming increasingly capable at tasks relevant to biological weapons research and cybersecurity operations (Mak & Götting, 2026; UK AI Security Institute, 2026). Recent incidents show that models are highly capable of performing complex hacking tasks in the real world (METR, 2026; OpenAI, 2026). Meanwhile, bad actors have attempted a wide range of misuse of frontier language models, including biological weapons research and bomb manufacturing (Anthropic, 2026b; Juelich, 2026).

The closed-weights nature of the current frontier models enables relatively strong defense strategies against misuse, such as refusal training and safety classifiers (Anthropic, 2026a), and robustly evading them requires significant effort (Zou et al., 2025). Open-weights models, however, trail close to the frontier closed-weights models' capabilities (UK AI Security Institute and U.S. Center for AI Standards and Innovation, 2026; Mak & Götting, 2026) without their safety measures. The same defense strategies do not work with open-weights models (Casper et al., 2026), due to the possibility of setting up private inference servers without safety classifiers and the ability to fine-tune models, a small amount of which can reverse standard safety training (Qi et al., 2024; Yang et al., 2024; Lermen et al., 2024; Zhan et al., 2024).

To address open-weights safety, we need models that are resistant to fine-tuning towards harmful capabilities (are "tamper-resistant"). Recent work has studied unlearning methods (applied after pretraining) to make it difficult to fine-tune models towards harmful capabilities (Rosati et al., 2024a; Tamirisa et al., 2025; Fan et al., 2025a), but their resistance to fine-tuning may break down under different fine-tuning conditions (Qi et al., 2025; Rosati et al., 2026). This implies that such methods do not robustly remove model capabilities. Meanwhile, the state-of-the-art approach for tamper resistance is to filter out harmful data in the pretraining phase (O'Brien et al., 2026). However, data filtering runs into steep diminishing returns once the most harmful data has already been removed.

Instead of trying to remove capabilities post-hoc, we apply unlearning algorithms during pretraining to prevent them from forming in the first place. Our experiments indicate that pretrain-time unlearning also composes with data filtering, and enables actively training against harmful capabilities instead of only passively filtering harmful data out. Figure 1 conceptualizes the overall approach.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Regular training (b) Data filtering (c) Pretrain-time unlearning

Figure 1: Data filtering improves tamper resistance, but the model may still generalize to the vicinity of harmful data even after filtering the most harmful data. In addition to data filtering, we apply explicit negative training to actively prevent harmful capabilities from forming.

We show that:

1. Applying unlearning during pretraining can improve tamper resistance significantly more than post-hoc unlearning.
2. Pretrain-time unlearning can improve tamper resistance after filtering most harmful data out, and results in better safety-capability tradeoffs than more aggressive filtering.
3. Weighted gradient ascent is a simple and effective pretrain-time unlearning algorithm with stable pretraining dynamics.
4. The tamper resistance improvements hold across an order of magnitude of compute scaling.

2 Related Work

Tamper resistance. The early unlearning literature was mostly focused on privacy and removing the influence of particular data points to the resulting model (Ginart et al., 2019; Bourtoule et al., 2021). Thus, resistance against fine-tuning was not a common concern. Adversarial fine-tuning was first proposed as an unlearning metric in Golatkar et al. (2020a), and was also included in Golatkar et al. (2020b) and Tarun et al. (2024), although the motivation was mostly as an auxiliary metric to measure the amount of information left in the weights. Tamper resistance to prevent misuse of foundation models with dual-use capabilities was first proposed in Henderson et al. (2023), and fine-tuning resistance was proposed as the main metric. It was also noticed that standard safety tuning of LLMs can be easily reversed with small amount of fine-tuning (Qi et al., 2024; Yang et al., 2024; Lermen et al., 2024; Zhan et al., 2024). Lynch et al. (2024) proposed evaluation metrics for unlearning methods for LLMs, and argued for adversarial fine-tuning. Rosati et al. (2024b) formalizes conditions for an effective defense against fine-tuning attacks. Later unlearning papers have reported fine-tuning resistance as a metric (Sheshadri et al., 2024), although many works have shown that most methods are not resistant against fine-tuning (Deeb & Roger, 2024; Hu et al., 2025; Łucki et al., 2025; Che et al., 2025). Some methods target explicitly improving fine-tuning resistance (Rosati et al., 2024a; Tamirisa et al., 2025; Fan et al., 2025a), but later evaluations have found claims of fine-tuning resistance to be significantly weaker under different fine-tuning conditions (Qi et al., 2025; Rosati et al., 2026).

Data filtering and gradient routing. The state-of-the-art approach to improve tamper resistance in LLMs is filtering out data related to the harmful capabilities during pretraining. In particular, O'Brien et al. (2026) trained 6.9B parameter language models and showed adversarial fine-tuning resistance of up to 300M tokens of biothreat-related text. Later work has explored filtering on the token-level instead of document-level (Rathi & Radford, 2026). An alternate approach to data filtering is localizing harmful data updates to specific parameters via gradient routing (Cloud et al., 2024). Shilov et al. (2025) showed that this approach can result in an improved tradeoff between

the harmful data loss and general capabilities than standard filtering, especially when the harmful document classification is noisy.

Gradient-ascent-based unlearning. A common tool in unlearning is applying gradient descent with an inverted cross-entropy loss (Golatkar et al., 2020a; Jang et al., 2023), often titled *gradient ascent* (GA). Kurmanji et al. (2023) proposed to pair it with regular training on retain data, and called it NegGrad+. Maini et al. (2024) use a similar objective with regular training on retain data and gradient ascent on forget data, and called it *gradient difference*. An issue with gradient ascent and gradient difference, however, is the unbounded ascent loss. Zhang et al. (2024) proposed an improved method, negative preference optimization (NPO), where the gradient is an adaptive weighting of GA such that the loss remains bounded. Fan et al. (2025b) point out that NPO has biases caused by the use of a reference model in the loss, and propose a simplified method, SimNPO, that improves unlearning. Wang et al. (2025) also simplify NPO and propose weighted gradient ascent (WGA), which applies different weightings on different tokens.

Non-transferable learning. The field of non-transferable learning has similar motivations to our method in deliberately avoiding generalization to specific datasets, with interventions often applied during pretraining. The original motivation was intellectual property protection (Wang et al., 2022), and many works aiming to prevent unwanted generalization followed (Wang et al., 2023a;b; Hong et al., 2024; Wang et al., 2024; Zhou et al., 2024; Deng et al., 2024; Peng et al., 2024). Many in the literature even use gradient ascent style losses for preventing generalization (Wang et al., 2022; 2023b; Peng et al., 2024). Later work has explicitly evaluated NTL methods for resistance to fine-tuning on the unwanted domain, with limited success, and has focused mostly on vision models (Hong et al., 2025).

3 METHODS

Assume that the dataset D has been split into a harmful set D_h and benign set D_b with some classification process. Note that even though D_b is called "benign", in practice the split to harmful and benign involves choosing some arbitrary classification threshold with a potentially inaccurate classifier. Thus, some data in D_b can generalize to harmful capabilities.

Data filtering. The data filtering intervention simply involves pretraining the model only on D_b. By contrast, a "baseline" or "unfiltered" model is trained on D. More extensive data filtering involves sorting the data in D_b with the classifier, and removing some percentage of the most harmful data according to the classifier.

Pretrain-time unlearning. When doing pretrain-time unlearning, an unlearning loss based on D_h is used at some frequency f of steps, spread uniformly throughout training. This can be studied in isolation (where the model is trained on D) or as an intervention combined with data filtering (where the model is trained on D_b, or a subset of it if using more extensive filtering). In practical terms, unlearning naturally composes with data filtering because $D_h = D - D_b$, making productive use of the data that would otherwise be filtered out. Unlearning steps and regular steps share the same learning rate schedule and optimizer variables, such as momentum.

Post-hoc unlearning. Similarly to pretrain-time unlearning, post-hoc unlearning uses an unlearning loss based on D_h, except the unlearning steps are only done at the end of training. It can also be studied either in isolation or combined with data filtering. Since it happens after pretraining, it uses a separate learning rate schedule and optimizer instance.

3.1 UNLEARNING METHODS

Our goal is to find a stable method to demonstrate the feasibility of pretrain-time unlearning. We focus on gradient ascent methods, as they are a popular family of loss functions for unlearning and easy to integrate in LLM pretraining stacks via only changing the next token prediction loss function.

Gradient ascent. The gradient ascent (GA) loss is

$$\mathcal{L}_{\text{GA}} = \mathbb{E}_{\mathbf{x}\sim D_h} \sum_{i=2}^{|\mathbf{x}|} \log p_\theta(\mathbf{x}_i|\mathbf{x}_{<i}), \tag{1}$$

where $p_\theta(\mathbf{x}_i|\mathbf{x}_{<i})$ is the next-token prediction probability assigned by the model. The loss is the negation of the standard cross-entropy. The intuition for gradient ascent is that it "reverses" the effect of training on some data point. The intuition becomes stronger when applied during pretraining instead of post-hoc, as the reversal happens significantly closer to the corresponding gradient descent step. However, an issue with gradient ascent is that the loss is unbounded.

Weighted gradient ascent. The weighted gradient ascent (WGA) loss (Wang et al., 2025) is

$$\mathcal{L}_{\text{WGA}} = \mathbb{E}_{\mathbf{x}\sim D_h} \sum_{i=2}^{|\mathbf{x}|} \text{stopgrad}\left(p_\theta(\mathbf{x}_i|\mathbf{x}_{<i})^\alpha\right) \log p_\theta(\mathbf{x}_i|\mathbf{x}_{<i}). \tag{2}$$

It includes a weighting on each token such that the loss on already incorrectly predicted tokens are attenuated. This avoids the loss growing without bound. The stopgrad ensures that the gradient contributions from individual tokens are regular cross-entropy based gradients. $\alpha = 0.0$ is equivalent to gradient ascent and higher values cause more attenuation. We use $\alpha = 1.0$ in our experiments.

Benign loss mix-in. To increase stability while unlearning, we mix in a matched amount of data from the benign dataset and apply the regular next-token prediction loss on it, making the full loss

$$\mathcal{L}_{\text{Full}} = \lambda\mathcal{L}_{\text{Unlearn}} + \mathcal{L}_{\text{Benign}}, \tag{3}$$

where $\mathcal{L}_{\text{Benign}}$ is the standard cross-entropy loss and λ is a hyperparameter trading off the strength of the unlearning gradient with the next-token prediction gradient. Note that the mixing procedure makes gradient ascent identical to gradient difference.

4 Experiments

In Section 4.1, we describe the broad experimental setup, the evaluation methodology, and the dataset used. In Section 4.2, we demonstrate that weighted gradient ascent provides much stabler pretraining dynamics than gradient ascent. In Section 4.3, we compare pretrain-time unlearning to the post-hoc unlearning and filtering baselines and find that it gives a consistent improvement across scales. In Section 4.4, we show that pretrain-time unlearning can provide better tradeoffs between tamper resistance and general capabilities than simply doing more aggressive data filtering.

4.1 Experiment setup

Tamper resistance metric: We evaluate the tamper resistance of the model as the difficulty of fine-tuning the model to improved performance on the harmful data. We consider two kinds of fine-tuning attacks: 1) compute-limited attacks, in which we assume infinite data and rank methods based on tokens of fine-tuning for the loss to reach a threshold, and 2) data-limited attacks, in which we assume infinite compute and rank methods on minimum achievable loss. With compute-limited attacks, the loss threshold is the harmful data loss of an unfiltered model trained with the same compute. For plots with runs from different compute scales and benign loss on the x-axis, we instead use the harmful loss of a model with the same benign loss, to make points with the same x-axis value directly comparable.

Data and classification: We focus the experiments on an English Wikipedia dataset (Wikimedia, 2025), consisting of 6.4 million articles and 3.7 billion total tokens. Following Shilov et al. (2025), we use a document-link based fastText classifier (Joulin et al., 2017) provided by Wikipedia's content classification service (Wikimedia, 2025) and define the harmful dataset as the documents with the highest "STEM.Biology" probability. This accounts for roughly 4% of the full dataset. We also use the classifier to determine benign set document similarity to Biology in Section 4.4.

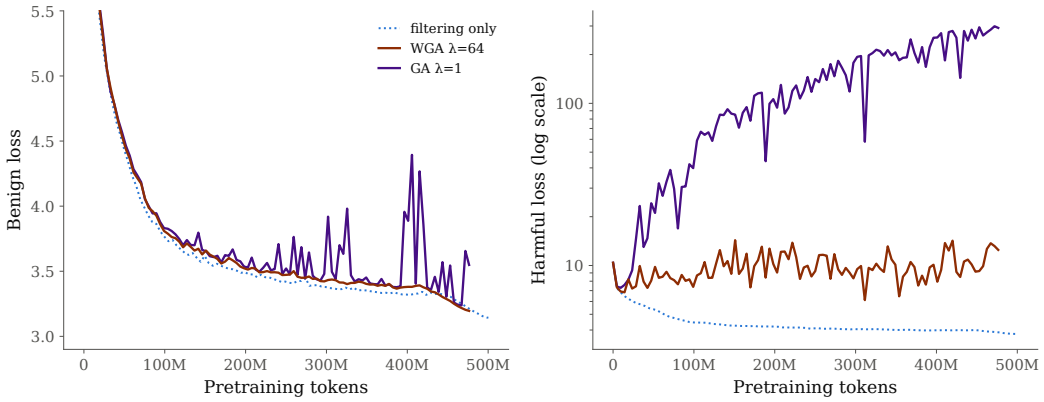


Figure 2: Left: WGA results in stabler learning dynamics than gradient ascent. Right: Gradient ascent has unbounded loss on the harmful data set, whereas WGA equilibrates at a maximum value.

Unlearning hyperparameters: Our plots use the weighted gradient ascent (WGA) unlearning method, applied over $f = 0.05$ of training steps. Section 4.2 shows the rationale behind WGA while Section C discusses f. $\alpha = 1.0$ is a default that works well and wasn't explicitly tuned. λ varies across methods and plots as it is the main knob to trade off benign loss and harmful fine-tuning resistance. All methods for a single model parameter scale are "token-matched", meaning they are trained with the same total number of tokens, counting both benign and harmful tokens. This approximately matches compute usage and is easier to define than compute. For pretrain-time unlearning, this means it uses 95% of the total training steps of filtering, because each unlearning step uses 2x the tokens of a normal step. Post-hoc unlearning uses 90% of the steps of filtering in pretraining, and then a 50%/50% mix of benign and harmful data in the unlearning phase, accounting for a total 10% of the compute. This matches unlearning tokens for pretraining and post-hoc unlearning.

Pretraining hyperparameters: We pretrain Transformers with architecture roughly matching Karpathy (2025), up to 163M parameters. Learning rate and batch size are set per scale with the methodology described in Li et al. (2025), and the # of train tokens is set to be compute-optimal according to Hoffmann et al. (2022). We include more details in Section A.

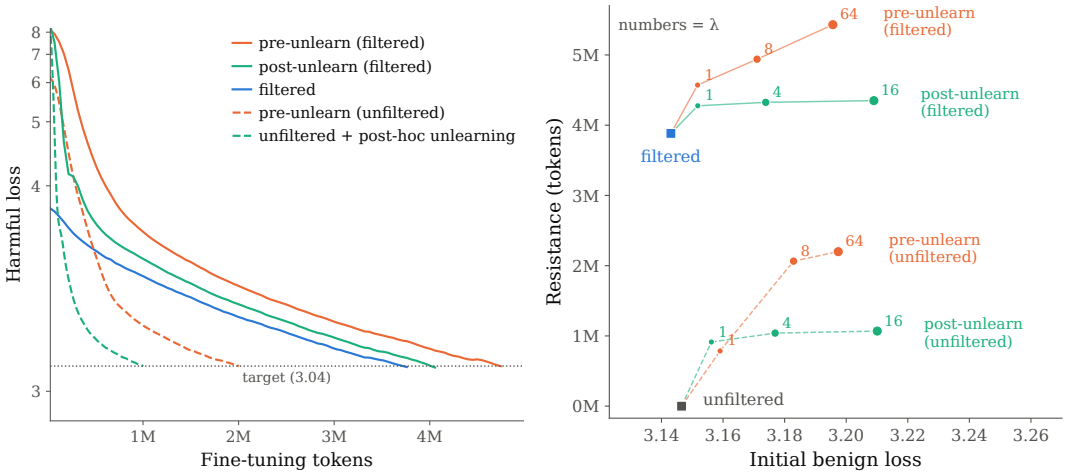
Adversarial fine-tuning hyperparameters: In the scaling analysis, for each scale and for each tamper resistance method, we separately tune the learning rate and batch size of the adversarial fine-tuning attack. The optimal fine-tuning learning rate and batch size can be found using a similar methodology as for pretraining, which we describe further in Section B. In the single-scale results, we use a lower-compute evaluation: We sweep over the learning rate for a single pretraining WGA run, and use the same learning rate for evaluating the other runs. This ensures that we do not choose the learning rate to be advantageous for the pretrain unlearning models.

4.2 Weighted gradient ascent has stable training dynamics

In Fig. 2, applying weighted gradient ascent (WGA) during pretraining results in overall stable pretraining dynamics, especially for benign loss. This is in contrast to standard gradient ascent (GA), which has unbounded harmful loss and tends to cause much more benign loss damage than WGA. GA starts creating significant benign loss spikes at $\lambda = 1.0$ while WGA causes no perceptible benign loss spikes even at $\lambda = 64.0$. Furthermore, GA increases harmful loss by an order of magnitude more than WGA and doesn't seem to saturate the way WGA does.

4.3 Unlearning is better done during pretraining

In these experiments, we compare running an unlearning algorithm during pretraining against running the same algorithm post-hoc for the same # of unlearning tokens. We evaluate both the setting where unlearning is the sole tamper resistance intervention and the setting where unlearning is com-



(a) Harmful loss curves during adversarial fine-tuning. Pre-unlearn methods use $\lambda = 8.0$ while post-unlearn methods use $\lambda = 4.0$.

(b) The tradeoff between fine-tuning resistance and benign loss for different values of λ.

Figure 3: Plots comparing pretrain-time unlearning, post-hoc unlearning, and filtering for a 36M parameter model. Resistance is defined as the # of tokens needed to reach the harmful loss of a 36M parameter unfiltered model.

posed together with filtering. Here, we focus on compute-limited attacks to illustrate the differences between post-hoc and pretrain-time unlearning.

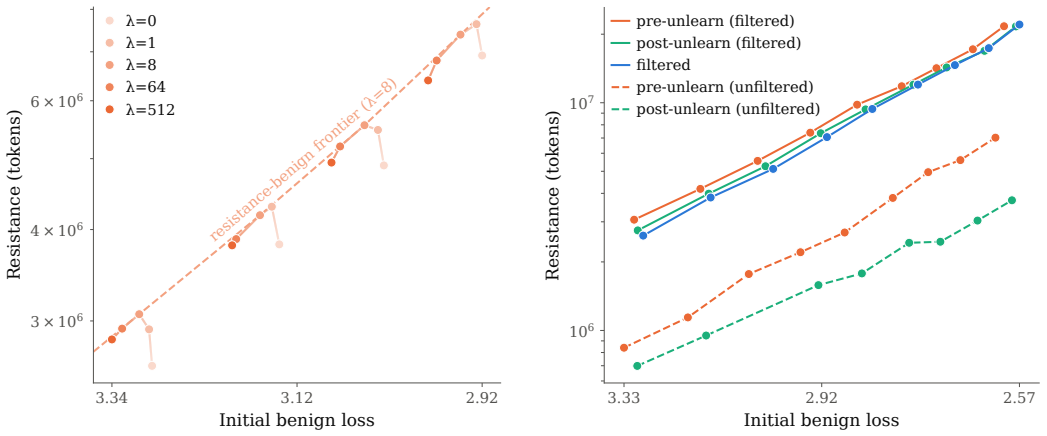
Adversarial fine-tuning example. In Fig. 3a, we show that unlearning during pretraining can enable much more tamper resistance than unlearning after pretraining. This holds both when unlearning is applied with and without filtering the harmful set from pretraining. Note that all of the unlearned models have a similar initial harmful loss, and the underlying resistance is only revealed via fine-tuning. The λ values are selected so that all unlearned models have similar initial benign losses.

Tradeoff between tamper resistance and capability. In Fig. 3b, we show the full tradeoff for each of these models for different values of λ. Improvements in fine-tuning resistance plateau quickly for the post-hoc unlearning methods, even though the benign loss increases. Pretrain-time unlearning, on the contrary, steadily increases the fine-tuning resistance when applied with or without data filtering. Note that in these plots, the same target loss is used for all points when defining fine-tuning resistance, which matches the invariant described in Section 4.1.

Resistance-benign frontier across scales. In Fig. 4a, we show that the same value of λ has a consistent effect on the resistance-benign tradeoff across scales. The frontier is composed of the λ values such that curve dominates all others at equal benign loss. Here, resistance is resistance is defined w.r.t. the harmful loss of a baseline model that reaches the same benign loss of the run (as extrapolated by a loss scaling law). This smoothly interpolates the target harmful loss in regions between different scales.

Since increases in λ increase benign loss, they also increase the target harmful loss. This leads to a decrease in the resistance metric, and shows up as a characteristic downward curve. In other words, past a certain value of λ, it becomes preferable to train a smaller model with a smaller λ to get higher resistance at the same benign loss. This λ forms the *resistance-benign frontier* and differs for pretrain-time and post-hoc unlearning. Fig. 4a shows that this value is $\lambda = 8.0$ for pretrain-time unlearning, while Section C.2 shows that it is $\lambda = 0.25$ for post-hoc unlearning.

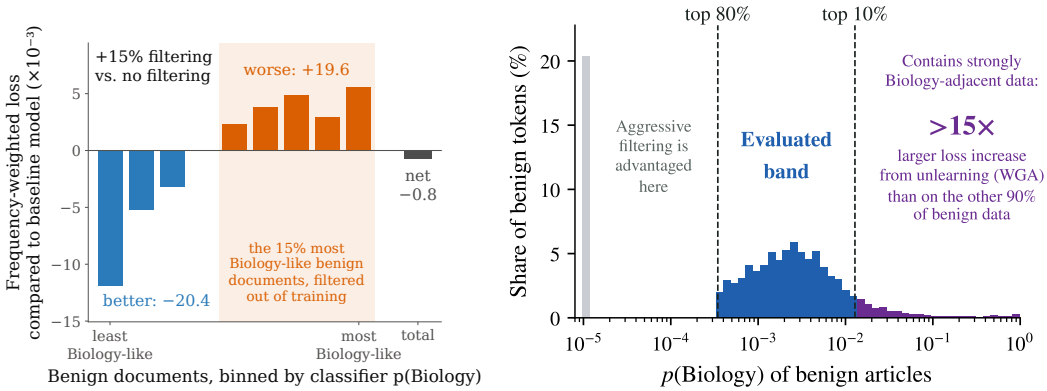
Scaling laws of tamper resistance. Fig. 4b compares the resistance-benign frontiers of different methods, finding that pretrain-time unlearning improves tamper resistance significantly more than post-hoc unlearning over unfiltered models, and the gap increases with scale. This effect persists when both unlearning methods are combined with data filtering as well. Additionally, pretrain-



(a) The resistance-benign tradeoffs of different values of λ for pretrain-time unlearning, from 27M to 58M parameters. Interpolating $\lambda = 8.0$ points forms the resistance-benign frontier.

(b) Resistance-benign frontiers of all methods from 27M to 163M parameters. Preunlearn methods use $\lambda = 8.0$ while postunlearn methods use $\lambda = 0.25$.

Figure 4: Plots showing the scaling laws of tamper resistance relative to the initial benign loss before fine-tuning. Resistance is defined as the # of tokens needed to reach the harmful loss of a unfiltered model that matches the initial benign loss.



(a) Loss distribution with aggressive filtering.

(b) $p(\text{Biology})$ histogram on Wikipedia.

Figure 5: Considerations for the evaluation metric: With aggressive filtering, the loss increase on filtered-out text is compensated by a loss decrease on the remaining benign text due to more optimization on that subdistribution. Our first evaluation metric accounts for this and excessive unlearning method loss on Biological text in the benign set by focusing on middle-band $p(\text{Biology})$ documents. See Fig. 12 for analysis on unlearned model loss on the top 10% data. Note that below the 20% quantile, the fastText classifier gives the same minimum value to all documents.

time unlearning has a consistently higher frontier than data filtering across scales, while post-hoc unlearning has a higher frontier initially but its advantage over filtering shrinks with scale.

4.4 Pretrain-time unlearning composes with filtering

Increased filtering baseline. Our method exposes a parameter that allows us to trade off capabilities for tamper resistance. A strong family of baselines with a similar parameter is simply filtering out the pretraining data more aggressively by using the document-link based classifier used for constructing the original topic classification.

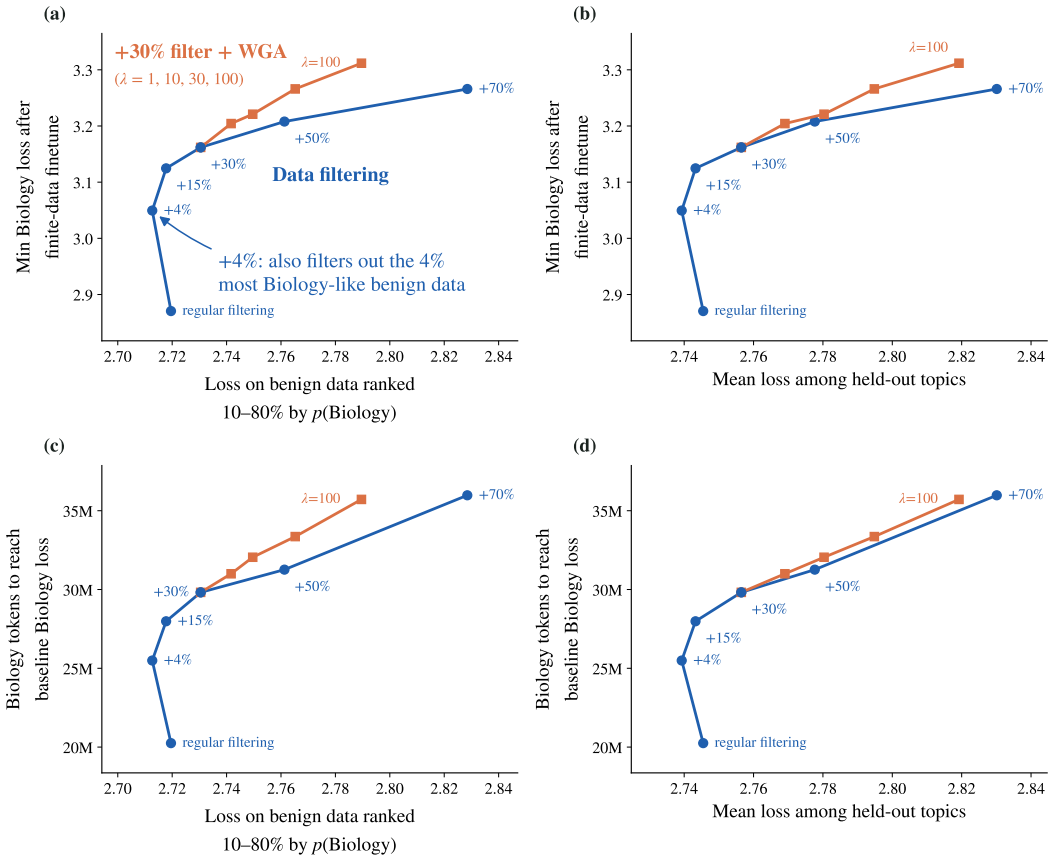


Figure 6: The safety-capability tradeoff for increasingly aggressive filtering and filtering + WGA, for different safety and capability metrics. Top row: The minimum achievable loss achievable with 0.52M Biology training data. Bottom row: Number of tokens until reaching a reference loss. Left: The mid-range Biology data loss, as ranked by the document classifier. Right: Loss on a set of topics held out from training, measuring generalization. All models are in the 163M parameter scale.

Why standard benign loss is not the right metric against filtering. On the full benign dataset loss, unlearning does not improve over more aggressive filtering (Section E, Fig. 14). However, we argue that the benign loss hides the full picture when comparing with pretrain-time unlearning, and should not be the main comparison metric: 1) The classification to the harmful and benign sets is imperfect, and the benign set includes some clearly Biological data (see Figs. 12 and 13 in Section E). Penalizing for a high loss on these would penalize the capability we are trying to remove. Unlearning methods are trained to explicitly increase loss on Biological documents, and any Biological documents in the benign set have high loss values that degrade the average significantly. 2) Aggressive filtering enables the training run to decrease loss on the specific subset of data that would be filtered last, offsetting the loss increase from filtered-out parts. This is visualized in Fig. 5a.

To measure performance outside the specific data distribution that aggressive filtering is trained on, and showcase the specific benefits of pretrain-time unlearning, we define two alternative capability metrics: 1) the loss on middle-band Biological documents, visualized in Fig. 5b, 2) the loss on held-out topics that are not trained on, measuring generalization outside the immediate data distribution. We also filter out the documents with the highest Biology scores from this set for the same Biological data outlier loss reason. We include more details about the choice of the held-out set and the results for the full benign set in Section E.

Unlearning improves over more filtering after filtering most harmful-adjacent data. We explore how well we can stack unlearning with data filtering in Fig. 6. To provide evidence for the benefit of pretrain-time unlearning, we evaluate both the compute-limited and data-limited metrics,

against both capability metrics. At early levels of additional filtering, we mostly remove genuinely biology-relevant data, and the resistance against fine-tuning increases sharply. The curves quickly flatten after filtering about 10% of the full dataset, however, likely due to a lack of biology-relevant data to filter out (see Fig. 5b). Applying unlearning after already filtering out, e.g., 30% of the data, allows us to increase resistance against fine-tuning even further, with smaller degradation in the capability metrics.

5 Conclusion

We have provided evidence that pretrain-time unlearning is an effective tool for improving tamper resistance over filtering, providing a productive use for the harmful data that would otherwise be unused. We have demonstrated that unlearning during pretraining is more effective than unlearning post-hoc, and that weighted gradient ascent is a simple method that results in stable pretraining dynamics. We have also shown that pretrain-time unlearning can compose with data filtering. While more aggressive filtering can be highly effective while there is more harmful data to filter out, the increase in tamper resistance slows down after a point, and unlearning can provide further tamper resistance from there with less capability degradation.

We believe pretrain-time unlearning shows promise for improving the security of production open-weights models. Future work should evaluate whether our scaling laws continue to hold to models large enough to directly measure harmful capability instead of using harmful loss as a proxy. Evaluating the method on larger and more diverse datasets would also be important. Another direction is improving the capability-safety tradeoff of pretrain-time unlearning further by using more advanced unlearning methods or applying unlearning losses only on a subset of tokens. Understanding the mechanisms through which pretrain-time unlearning results in higher tamper resistance more deeply could provide high-level insights into such improvements.

AI use statement

In this work, we used generative AI tools for coding assistance, implementing methods, searching for relevant literature, giving feedback on writing, and creating plots. We have not used generative AI tools to generate synthetic data sets, draft the paper's text, help develop theoretical models or conceptual frameworks, formulate mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Section 3 describes how unlearning is integrated into training and Section 3.1 describes the unlearning methods themselves. Section 4.1 defines the metrics used in the paper, describes the data preprocessing steps, and records the key hyperparameters for experiments. Section 4.3 and Section 4.4 provide additional detail for data processing, metrics, and hyperparameters relevant to specific experiments. Section A lists the pretraining hyperparameters at each scale, Section B provides detail on how the fine-tuning hyperparameters for each data point were found, and Section C gives information on more detailed unlearning hyperparameters.

References

Anthropic. System card: Claude Fable 5.1 & Claude Mythos 5.1. Technical report, Anthropic, September 2026a. URL https://www-cdn.anthropic.com/0339e6a7c5c7b87f5c07798616dc32c215d14235/Claude%20Fable%205.1%20&%20Claude%20Mythos%205.1%20System%20Card.pdf.

Anthropic. Detecting and countering misuse of AI: September 2026. Technical report, Anthropic, September 2026b. URL https://www.anthropic.com/

threat-intelligence-report-september-2026. PDF: https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf.

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

Stephen Casper, Kyle O'Brien, Shayne Longpre, Elizabeth Seger, Kevin Klyman, Rishi Bommasani, Aniruddha Nrusimha, Ilia Shumailov, Sören Mindermann, Steven Basart, et al. Open technical problems in open-weight ai model risk management. *arXiv preprint arXiv:2608.07514*, 2026.

Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. Model tampering attacks enable more rigorous evaluations of LLM capabilities. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=E60YbLnQd2.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv preprint arXiv:2410.04332*, 2024.

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

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2553–2571. IEEE, 2024.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *Advances in Neural Information Processing Systems*, 38:1540–1567, 2025b.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

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)*, pp. 9301–9309. IEEE, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, pp. 383–398. Springer, 2020b.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FYKVPOHCpE.

Ziming Hong, Yongli Xiang, and Tongliang Liu. Toward robust non-transferable learning: A survey and benchmark. *arXiv preprint arXiv:2502.13593*, 2025.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, April 2017.

Antonia Juelich. "God has helped us, and so will AI": How the terrorist group Boko Haram uses frontier AI. Frontier AI Working Paper 1/2026, Cambridge Programme on AI Science & Policy, University of Cambridge, July 2026. URL https://casp.ac/reports/ai-enabled-terrorism. PDF: https://casp.ac/__l5e/assets-v1/8e5796ad-c373-4dea-8470-466263cd2125/ai-enabled-terrorism-report.pdf.

Andrej Karpathy. nanochat: The best chatgpt that $100 can buy, 2025. URL https://github.com/karpathy/nanochat.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA fine-tuning efficiently undoes safety training in llama 2-chat 70b. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=Y52UbVhglu.

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Hanshan Zhang, Zili Wang, Shijie Xuyang, Yuantao Fan, Zhenyu Ding, Haoying Wang, Ning Ding, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale: Part i, step law – optimal hyperparameter scaling law in large language model pretraining, 2025. URL https://arxiv.org/abs/2503.04715.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=J5IRyTKZ9s.

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

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Nelly Mak and Jasper Götting. Introducing VCT-v2 — the updated Virology Capabilities Test. SecureBio blog, September 2026. URL https://securebio.org/blog/introducing-vct-v2/.

METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the openai / hugging face hacking incident. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/, 08 2026.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella R Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. In *International Conference on Learning Representations*, volume 2026, pp. 35579–35633, 2026.

OpenAI. The Hugging Face incident and the road ahead. OpenAI blog, August 2026. URL https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

Boyang Peng, Sanqing Qu, Yong Wu, Tianpei Zou, Lianghua He, Alois Knoll, Guang Chen, and Changjun Jiang. Map: Mask-pruning for source-free model intellectual property protection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23585–23594. IEEE, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight llms. In *International Conference on Learning Representations*, volume 2025, pp. 62651–62681, 2025.

Neil Rathi and Alec Radford. Shaping capabilities with token-level data filtering. *arXiv preprint arXiv:2601.21571*, 2026.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024a.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5234–5247, 2024b.

Domenic Rosati, Xijie Zeng, Hong Huang, Sebastian Dionicio, Subhabrata Majumdar, Frank Rudzicz, and Hassan Sajjad. Limits of convergence-rate control for open-weight safety. *arXiv preprint arXiv:2602.18868*, 2026.

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

Igor Shilov, Alex Cloud, Aryo Pradipta Gema, Jacob Goldman-Wetzler, Nina Panickssery, Henry Sleight, Erik Jones, and Cem Anil. Beyond data filtering: Knowledge localization for capability removal in llms. *arXiv preprint arXiv:2512.05648*, 2025.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. Tamper-resistant safeguards for open-weight llms. In *International Conference on Learning Representations*, volume 2025, pp. 101802–101829, 2025.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE transactions on neural networks and learning systems*, 35(9):13046–13055, 2024.

UK AI Security Institute. Our evaluation of Claude Mythos Preview's cyber capabilities. AI Security Institute blog, April 2026. URL https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

UK AI Security Institute and U.S. Center for AI Standards and Innovation. UK AISI / CAISI preliminary assessment of Kimi K3's cyber capabilities. AI Security Institute blog, July 2026. URL https://www.aisi.gov.uk/blog/preliminary-assessment-of-kimi-k3s-cyber-capabilities.

Haotian Wang, Haoang Chi, Wenjing Yang, Zhipeng Lin, Mingyang Geng, Long Lan, Jing Zhang, and Dacheng Tao. Domain specified optimization for deployment authorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5095–5105, October 2023a.

Lianyu Wang, Meng Wang, Daoqiang Zhang, and Huazhu Fu. Model barrier: A compact untransferable isolation domain for model intellectual property protection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20475–20484. IEEE, 2023b.

Lianyu Wang, Meng Wang, Huazhu Fu, and Daoqiang Zhang. Say no to freeloader: Protecting intellectual property of your deep model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11073–11086, 2024.

Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tYRrOdSnVUy.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q. Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. In *International Conference on Learning Representations*, volume 2025, pp. 61897–61931, 2025.

Wikimedia. Wikimedia downloads, 2025. URL https://dumps.wikimedia.org.

Wikimedia. Ores / articletopic. https://www.mediawiki.org/wiki/ORES/Articletopic, 2025. Accessed: 2025-09-21.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Ruth Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=9qymw6T9Oo.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori B Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 681–687, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Tong Zhou, Shaolei Ren, and Xiaolin Xu. Archlock: Locking DNN transferability at the architecture level with a zero-cost binary predictor. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=e2YOVTenU9.

Andy Zou, Maxwell Lin, Eliot Jones, Micha Nowak, Mateusz Dziemian, Nick Winter, Alexander Grattan, Valent Nathanael, Ayla Croft, Xander Davies, Jai Patel, Robert Kirk, Nate Burnikell, Yarin Gal, Dan Hendrycks, J. Zico Kolter, and Matt Fredrikson. Security challenges in ai agent deployment: Insights from a large scale public competition. *Advances in Neural Information Processing Systems*, 38, 2025.

A Pretraining Details

Table 1 lists the main hyperparameters at each scale used during pretraining. We use Muon for matrix parameters and AdamW for other parameters, sharing a learning rate between them based on Liu et al. (2025). We use a WSD learning rate schedule with 2% warmup and 10% decay. The learning rate and batch size are set with scaling curves based on the methodology described in Li et al. (2025). The # of train tokens is set to be compute-optimal with IsoFLOPs curves, and then adjusted for the unlearning methods as described in Section 4.1. All models use $d_{\text{vocab}} = 32768$, $n_{\text{ctx}} = 2048$, and weight decay 0.1.

# params	# layers	d_{model}	peak LR	batch size	# tokens
27M	5	320	3.7×10^{-3}	64	358M
36M	6	384	3.2×10^{-3}	64	500M
46M	7	448	2.9×10^{-3}	96	667M
58M	8	512	2.6×10^{-3}	96	858M
74M	9	576	2.3×10^{-3}	96	1.08B
92M	10	640	2.1×10^{-3}	128	1.33B
111M	11	704	2.0×10^{-3}	128	1.59B
135M	12	768	1.8×10^{-3}	128	1.90B
163M	13	832	1.7×10^{-3}	192	2.24B

Table 1: Pretraining hyperparameters.

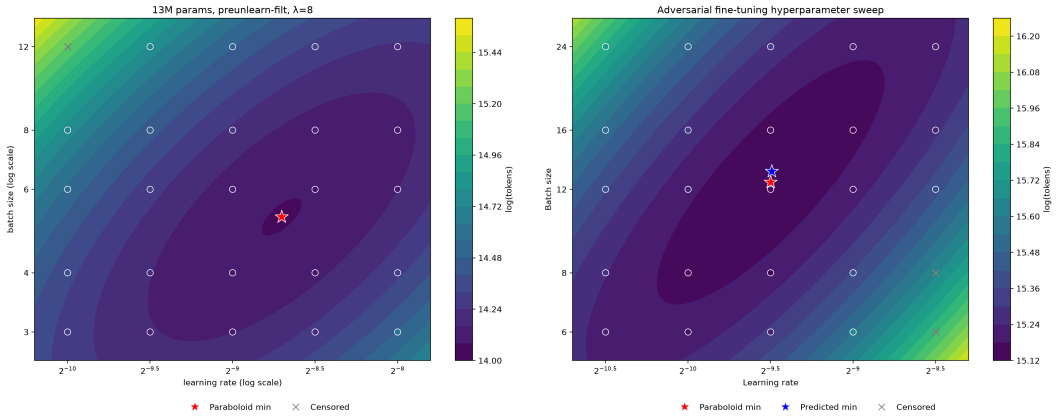


Figure 7: Paraboloid fits for the optimal learning rate and batch size during adversarial fine-tuning.

B Adversarial Fine-Tuning Details

We follow a similar procedure as in pretraining to tune the learning rate (η) and batch size (B) for the adversarial fine-tuning runs of each scale and tamper resistance method. For any individual run, $\log(\text{fine-tuning tokens})$ can be modeled as a paraboloid w.r.t. $\log(\eta)$ and $\log(B)$, as shown in Fig. 7. After fitting paraboloids at each scale, we fit a power law for each tamper resistance method to the optimal η and B as a function of model size, as shown in Fig. 8

C Unlearning Details

C.1 Strong unlearning substitutes for frequent unlearning

One could be concerned that mixing unlearning steps with regular pretraining steps increases the total compute, making pretrain-time unlearning overall uncompetitive. In Fig. 9, we show that we can trade off the frequency of applying unlearning steps f with the unlearning loss scaling λ surprisingly well, enabling using unlearning only for a small fraction of the total steps. Within reasonable bounds, the product $\lambda \times f$ predicts the steps to relearn the harmful topic as well as the general capabilities loss. Overall, choosing f between 5% and 100%, it is possible to achieve roughly the same (loss, resistance against fine-tuning) tradeoff by varying λ accordingly.

C.2 Resistance-benign frontier for post-hoc unlearning

The resistance-benign frontier for post-hoc unlearning is formed by $\lambda = 0.25$, as shown in Fig. 10. This is much lower than the value for pretrain-time unlearning, which is $\lambda = 8.0$ (see Figure 4a).

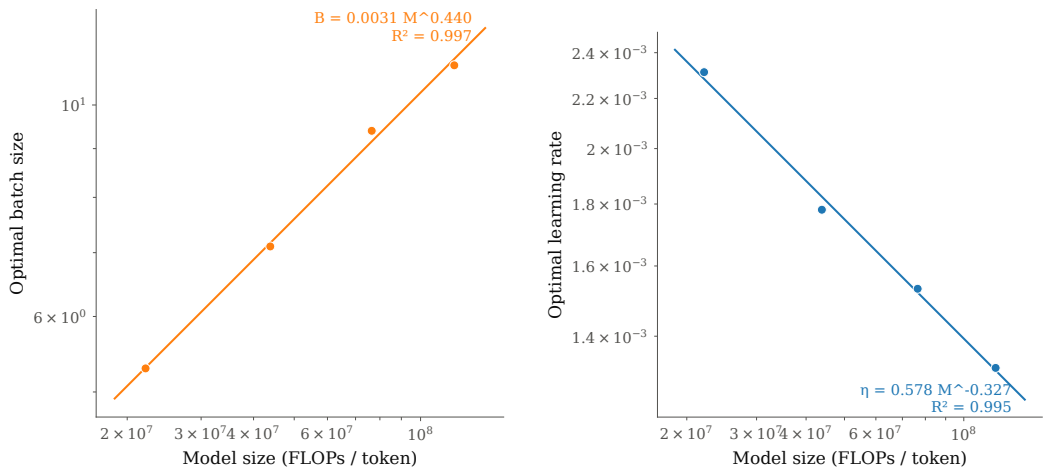


Figure 8: Power law fits for the optimal learning rate and batch size during adversarial fine-tuning across scales.

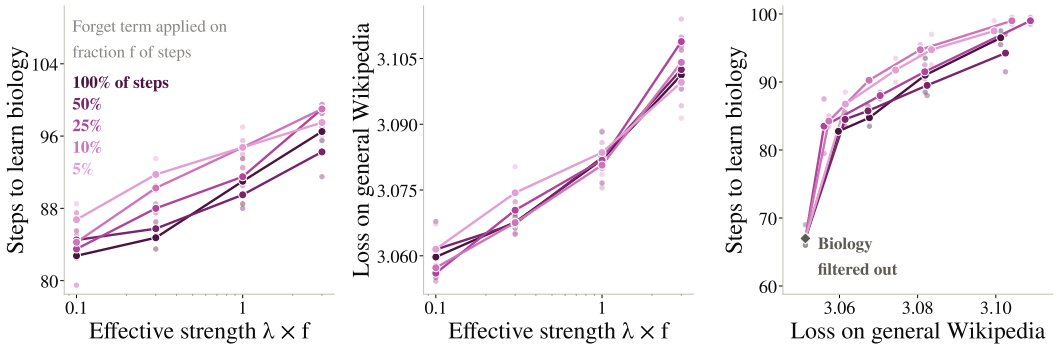


Figure 9: The interaction with the unlearning loss weight λ and the frequency of applying unlearning steps f with standard WGA. $\lambda \times f$ is a good predictor of fine-tuning resistance and general Wikipedia loss.

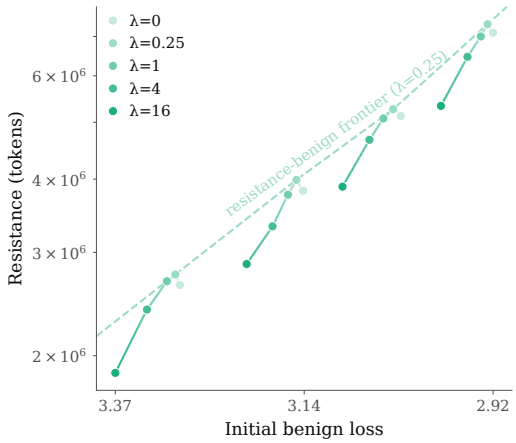


Figure 10: The resistance-benign tradeoffs of different values of λ for post-hoc unlearning, from 27M to 58M parameters. Interpolating $\lambda = 0.25$ points forms the resistance-benign frontier.

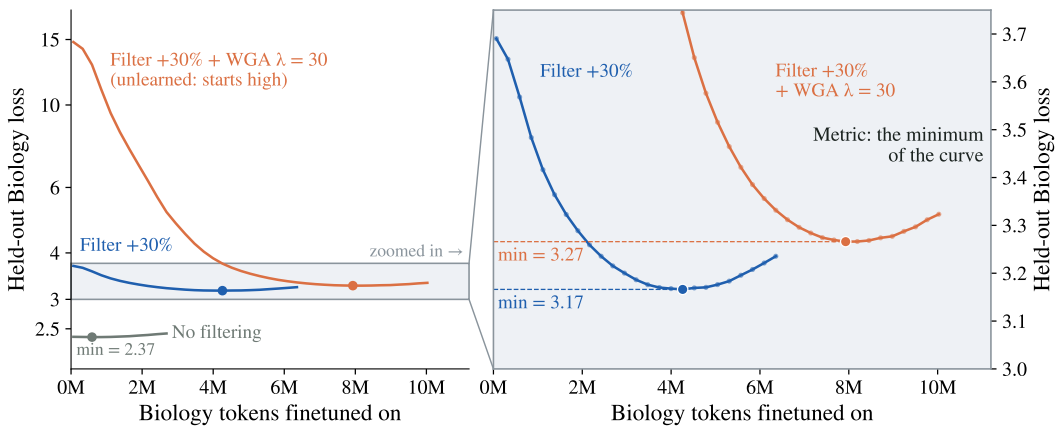


Figure 11: A visualisation of the finite-data fine-tuning metric, with data from the runs in Fig. 6. We choose a fixed set of 0.52M Biology topic tokens, and fine-tune until reaching an optimal validation loss.

D THE FINITE DATA FINE-TUNING EVALUATION METRIC

Figure 11 visualizes the finite-data fine-tuning evaluation. Using a fixed amount of Biology text (0.52M tokens from the Biology set), we fine-tune with multiple epochs until the validation loss starts increasing. This measures how well the model can generalize from finite data.

E ADDITIONAL DETAILS ON COMPOSING PRETRAIN-TIME UNLEARNING WITH FILTERING

This section contains additional details and evaluations for the experiments in Section 4.4. Figure 12 showcases the loss increase on the top-10% most Biology-like documents due to WGA, motivating the omission of those documents in the middle-band capability metric. Similarly, Fig. 13 shows example histograms of WGA-pretrained model NLLs on held-out documents in the held-out topic evaluation. Many of the topics contain documents that overlap heavily with Biology and appear as outliers in NLL, and filtering out the top-2% most Biology-like documents removes them.

Figure 14 shows the results in Fig. 6 for different capability metrics: WGA outperforms increased data filtering on loss outside the top-10% most Biology-like documents of the benign set, but not when considering the full benign set. This is likely due to the extreme concentration of loss in these highly Biology-like documents in unlearned models, as shown in Fig. 12.

Table 2 shows the held-out topics for the held-out topic loss evaluation in Fig. 6. Table 3 shows another held-out set, and Fig. 15 compares the results for the two held-out sets, showing that the results are similar across different splits. Both topic sets were obtained by randomly ordering the topics and adding the first topics in the sequence until they sum up to at least 9.5% of the total tokens in the data. The topics are rejected if the list consists of less than 6 topics, and they sum up to more than 10.5% of the data. In that case, the topics are reordered again and the procedure continues until we find a valid topic set. Some topics were excluded from consideration: 1) Biology due to it being the harmful set, and 2) Medicine and Health, Earth and Environment and Chemistry due to them being highly Biology-adjacent. Figure 16 showcases the topic composition with increased filtering of the benign set.

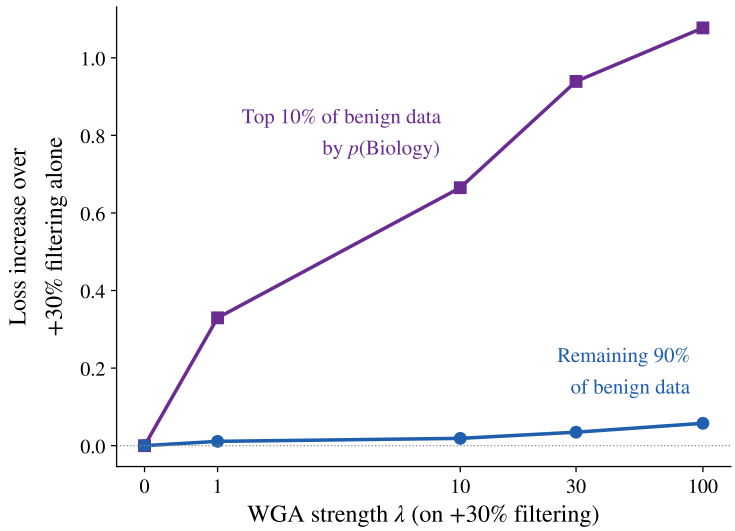
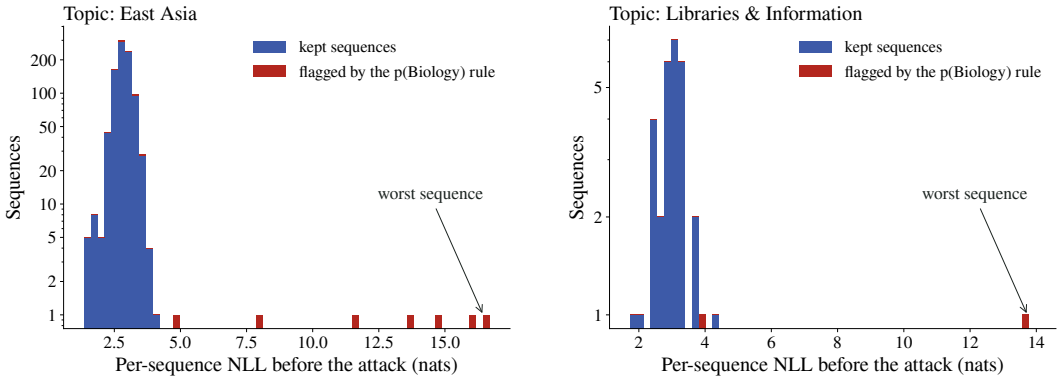


Figure 12: The loss increase on the top 10% most Biology-like documents due to WGA (Fig. 6), vs. the loss increase on the bottom 90%. The top 10% accounts for a disproportionate amount of the loss increase caused by WGA.



Worst sequence #766, mostly "Kikutaro Baba"

```
: Eubranchidae). Publications of the Seto Marine
   Biological Laboratory 19(1):39-42, pl. 1.
Baba, K. 1972. Thecacera picta spec. nov. from Suruga
   Bay, Japan (Nudibranchia: Doridoidea: Polyceridae).
   Veliger 15(2):88- 90.
Robilliard, G.A. & Baba, K. (1972). "Aldisa sanguinea
   cooperi subsp. nov. from the coast of the state of
   Washington, with notes on its feeding and spawning
   habits (Nudibranchia, Doridiidae, Aldisinae)". Pub.
   Seto Marine Biol. Lab., 19(6, March): 409-414.
Baba, K. 1974. Notes on Lobiger souverbii Fischer, 1856,
   re-identified, of Japan (Opisthobranchia: Sacoglossa:
   Lobigeridae). Collecting & Breeding 36(4): 74-77.
Baba, K. 1974. Some Comments on Lobiger-Souverbii Re
   Identified of Japan Opisthobranchia Sacoglossa
   Lobigeridae. Veliger 16 (3). 1974. 253-257.
Baba, K. 1974. List of the species of Phestilla from the
   central and western Pacific. Chiribotan (Newsletter
   of the Malacological Society of Japan) 8(3): 51-52.
Baba, K. 1974. New distributional record of Aegires
```

Worst sequence #4, mostly "List of MeSH codes (D12.776.476)"

```
kinase 4
- cyclin-dependent kinase 6
- maturation-promoting factor
- cdc2 protein kinase
- eif-2 kinase
- focal adhesion protein-tyrosine kinases
- glycogen synthase kinases
- glycogen synthase kinase 3
- gtp-binding protein regulators
- gtpase-activating proteins
- chimerin proteins
- chimerin 1
- eukaryotic initiation factor-5
- ras gtpase-activating proteins
- neurofibromin 1
- p120 gtpase activating protein
- rgs proteins
- guanine nucleotide dissociation inhibitors
- guanine nucleotide exchange factors
- eukaryotic initiation factor-2b
```

Figure 13: With the held-out topic evaluation, some documents have extreme NLL values with the WGA-pretrains. On closer inspection, these are highly Biology-relevant pieces of text, and they can be filtered out with the Biology classifier.

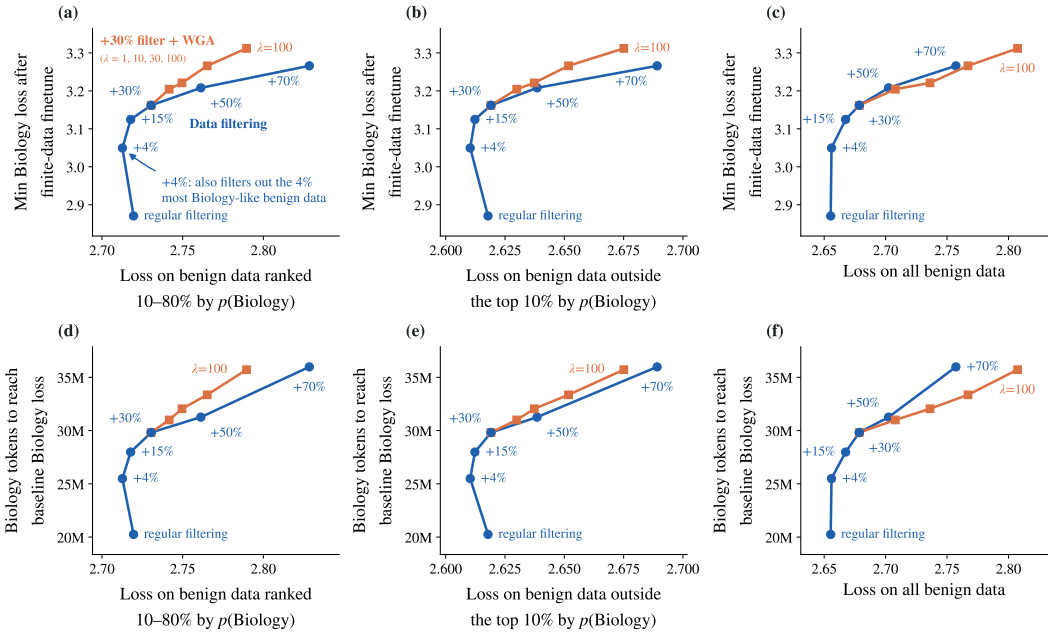


Figure 14: The safety-capability comparison following the experimental procedure in Fig. 6, with different subsets of the benign data on the x-axis. Unlearning improves over more aggressive filtering as long as we don't consider the top 10% most Biology-like data in the benign set.

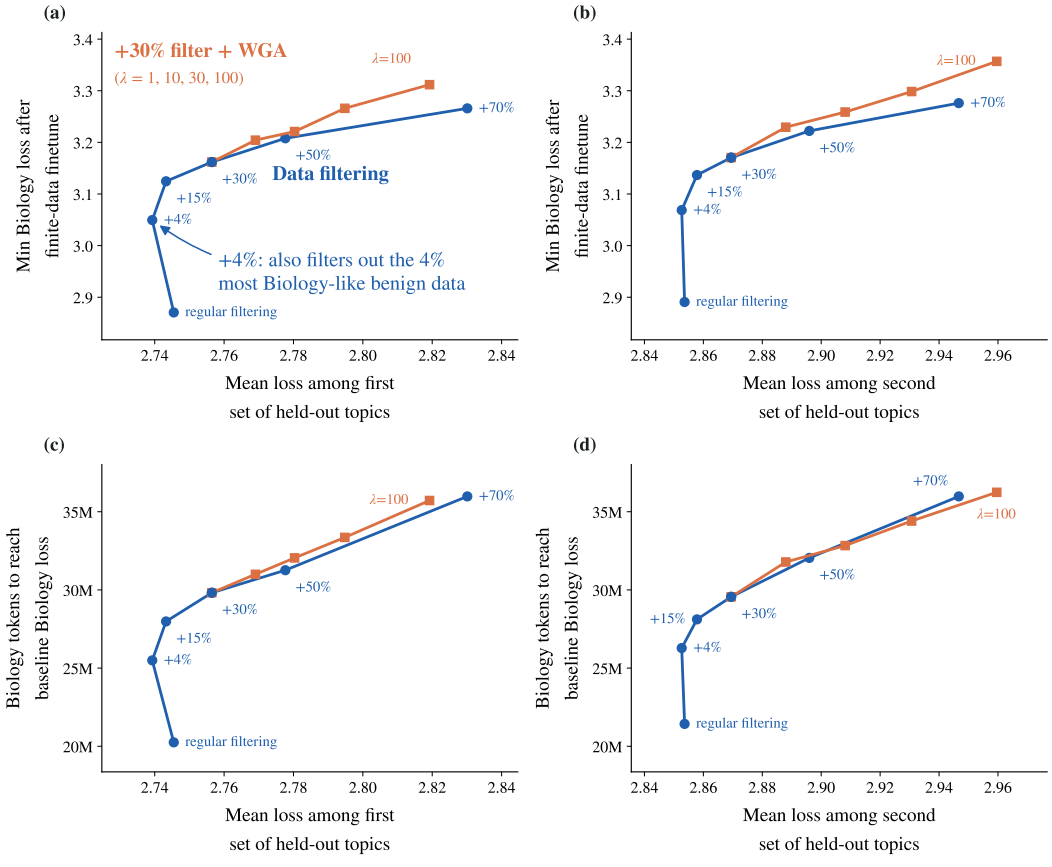


Figure 15: Comparison between different held-out splits, for the experimental setup in Fig. 6. The topics used are listed in Table 2 and Table 3.

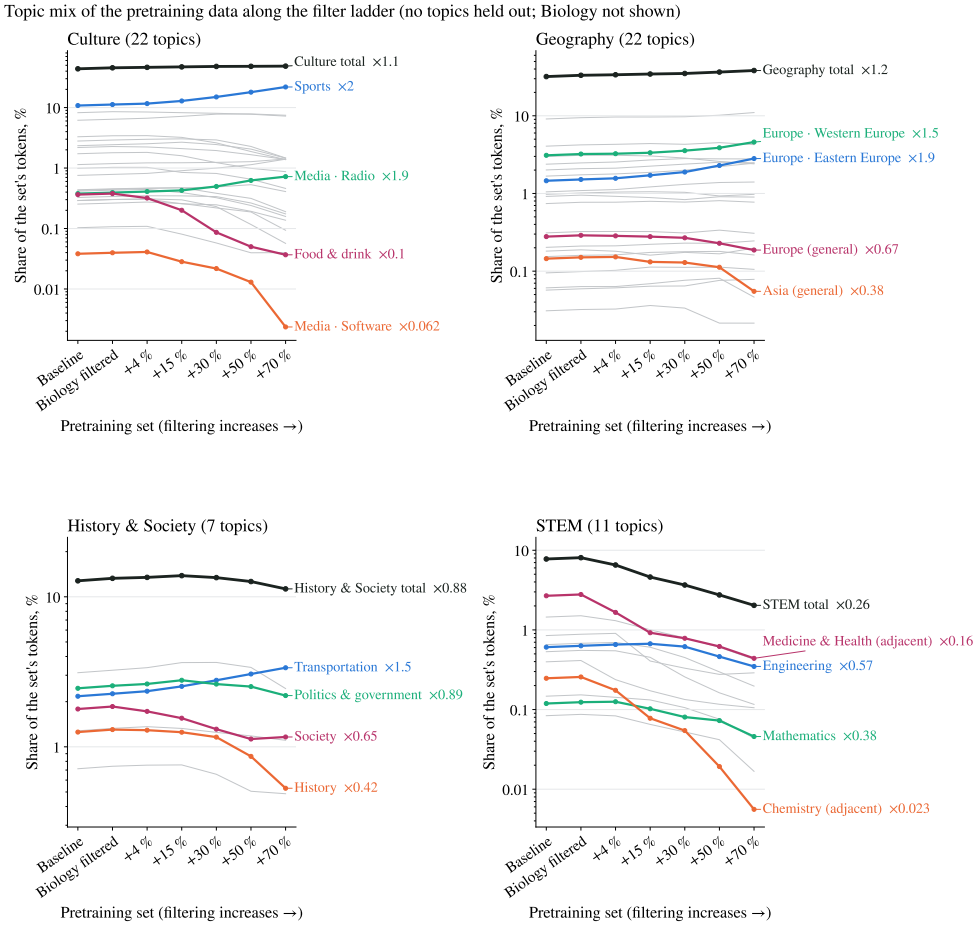


Figure 16: Topic composition in the benign set after aggressive filtering, with some topics highlighted.

Table 2: The first set of held-out topics: 8 ORES topics, 9.95% of validation tokens.

Category	Topic	Share of tokens
Geography	Western Europe	3.10%
History & Society	Politics and government	2.46%
History & Society	Transportation	2.17%
STEM	Space	0.85%
Culture	Entertainment	0.41%
Culture	Radio	0.38%
Culture	Media (general)	0.33%
Culture	Comics and Anime	0.25%
Total		9.95%

Table 3: The second set of held-out topics: 11 ORES topics, 9.76% of validation tokens.

Category	Topic	Share of tokens
Geography	Northern Europe	4.06%
Geography	East Asia	2.01%
Geography	Geographical	0.92%
Geography	South America	0.74%
STEM	Technology	0.53%
Culture	Internet culture	0.43%
Culture	Entertainment	0.41%
Culture	Visual arts (general)	0.29%
Geography	Southern Africa	0.18%
Culture	Fashion	0.10%
STEM	Libraries & Information	0.08%
Total		9.76%