Severance: When Unlearning Crime Unlearns Survival

Hannah Levin
Department of Computer Science
Stanford University

Nils Kuhn
Department of Electrical Engineering
Stanford University

Herman Sahota
NVIDIA

1 Introduction

Large language models acquire harmful capabilities as a side effect of pretraining on massive, largely unfiltered datasets. Post-training can suppress the surface expression of these capabilities. However, they leave the underlying representations intact, making models vulnerable to relearning attacks in which an adversary recovers the forgotten behavior through a small amount of finetuning.

Lee et al. (2025) propose that distillation robustifies unlearning and show that transferring an unlearned model's behavior into a fresh parameter space can prevent latent capabilities from being recovered. Their experiments are conducted on custom-pretrained models using language and arithmetic tasks, as well as the WMDP benchmark (Li et al. (2024)).

In this work, we reproduce, critique, and extend their findings. We replicate the core UNDO result on Pythia-160M language model (Biderman et al. (2023)) using the Task of Fictitious Unlearning (TOFU) benchmark (Maini et al. (2024)), extending their findings to smaller models and a new data setting. We evaluate the quality of unlearning using several attack strategies such as weight quantization, ease of relearning and Membership Inference Attack (Shokri et al. (2016)). Within the UNDO framework, we additionally introduce refusal unlearning as a variant that trains the model to produce refusal responses rather than random tokens. We critically evaluate key assumptions of the approach, and extend the work by evaluating unlearned models in a multi-agent simulation environment, testing whether unlearning translates to meaningful behavioral change in a live agent setting.

2 Related Work

2.1 Dataset

The TOFU benchmark Maini et al. (2024) provides a controlled setting for evaluating LLM unlearning by constructing a dataset of 4,000 question-answer pairs drawn from 200 synthetic author profiles generated by GPT-4. Because the authors are entirely fictitious, the exact source of the knowledge to be forgotten is known and controlled. This eliminates the ambiguity that arises when unlearning from real pretraining data, where the origin and extent of the learned information is rarely clear. Crucially, TOFU also enables a meaningful gold standard comparison, by making it possible to approximate what a model that never saw the forget data would look like. This property makes TOFU particularly well suited to our study.

2.2 LLM Unlearning

Machine unlearning refers to the problem of removing the influence of specific training data or capabilities from a trained model without retraining from scratch (Liu et al. (2025)). In the context of LLMs, this is particularly challenging because models trained on large, unfiltered datasets can memorize sensitive or harmful information, such that it is difficult to localize and remove. A central challenge is that existing unlearning methods tend to suppress unwanted capabilities without removing the underlying representations. Relearning attacks, in which an adversary deliberately finetunes the model on a small sample of the forgotten data, have repeatedly proven sufficient to recover suppressed capabilities (Lee et al. (2025)). More troublingly, even ordinary finetuning for unrelated tasks can cause suppressed behaviors to re-emerge (Liu et al. (2025)). This vulnerability motivates the search

for more robust unlearning methods that go beyond surface-level output modification, which is the central concern of the paper we engage with in this work.

2.3 Methods: Distillation and UNDO

Lee et al. (2025) address the problem of robust unlearning: producing a model that suppresses undesired capabilities and resists adversarial adversarial prompts, jailbreaks or finetuning. Their first key finding is that oracle matching is insufficient for robust unlearning. Even when a model perfectly replicates the output distribution of a model that never saw the forget data, the underlying latent capabilities remain intact and can be quickly recovered through finetuning. Their main result is that distillation robustifies unlearning. Training a randomly initialized student to match the outputs of an unlearned teacher transfers desired behavior without the training of undesired latent representations. Building on this, they introduce UNDO (Unlearn-Noise-Distill-on-Outputs), which avoids full reinitialization by instead corrupting the unlearned model's weights with noise:

$$\theta_{\text{perturbed}} = (1 - \alpha)\theta_{\text{suppressed}} + \alpha\beta N \tag{1}$$

α controls the degree of corruption and β scales the injected noise N. Distilling back into this noised model creates a tunable tradeoff between compute cost and robustness against relearning attacks.

3 Experiments

We evaluate unlearning methods on the TOFU benchmark Maini et al. (2024), splitting the dataset into a 90% retain set and a 10% forget set, corresponding to 180 authors to retain and 20 to forget. Since Pythia-160M Biderman et al. (2023) was not pretrained on TOFU, we first finetune it on the full dataset before applying any unlearning method. Each experiment follows a four-step pipeline: finetuning, unlearning, distillation, and a relearning attack. We evaluate unlearning robustness by measuring perplexity on the forget and retain sets before and after a relearning attack. During the relearning attack, the model is finetuned on the forget set for three epochs. A robust unlearning method should maintain high perplexity on the forget set throughout the relearning attack while preserving low perplexity on the retain set. The fresh Pythia model, which has never seen the TOFU data, serves as the gold standard. It is expected to maintain the highest forget perplexity throughout relearning.

3.1 Baseline Unlearning Methods

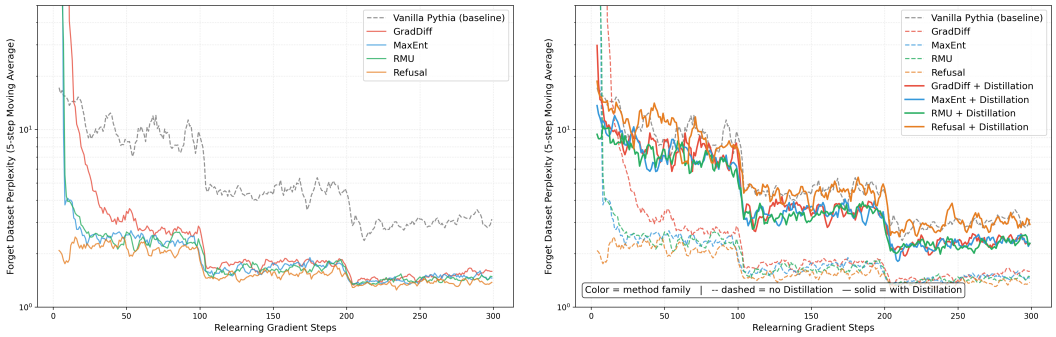
Table 1 reports perplexity after applying each unlearning method without distillation. Gradient Difference (GradDiff) Maini et al. (2024) maximizes the loss on the forget set while minimizing it on the retain set. Because it directly pushes token probabilities toward away from the optimum, the forget loss grows exponentially. This creates numerical instability under multi-epoch training and degrades retain performance as a side effect. Representation Misdirection for Unlearning (RMU) Li et al. (2024) operates by misdirecting internal representations at a specific layer. This produces strong forget perplexity but at a small cost to retain performance. Maximizing Entropy (MaxEnt) Liu et al. (2025) is the most stable of the three, achieving a favorable retain-forget tradeoff with less sensitivity to hyperparameters. As a fourth method, we introduce refusal unlearning, which trains the model to produce a refusal response such as "I cannot answer this question" rather than random or uninformative tokens. This preserves general language behavior more cleanly, as the model is never trained to produce incoherent outputs. Accordingly, refusal unlearning achieves the best retain perplexity of all methods, though it does not significantly reduce forget perplexity on its own.

Figure 2 (left) shows relearning curves for all baseline methods. Despite large differences in initial forget perplexity, all methods converge to similar relearning trajectories within one epoch, and none shows a meaningful advantage over the others by the third epoch. This confirms the central claim of Lee et al. Suppressing forget behavior does not prevent it from being recovered. A broader hyperparameter study both for unlearning methods, as well as for the retraining methods would be needed to identify an optimal method and optimal hyperparameters.

Since Pythia-160M was never trained on TOFU, it allows for an interesting separation between gold standard definitions. One gold standard can be defined as a model that has seen neither the forget nor the retain data. A second definition is a model that has seen the retain data but not the forget

Method	Perplexity on the forget set	Perplexity on the retain set	Unlearning Epochs
Fresh Pythia	26.79	26.33	0
Finetuned	2.16	2.23	0
GradDiff	$1.19 \cdot 10^8$	6.90	3
MaxEnt	$5.03 \cdot 10^4$	2.40	5
RMU	$2.00 \cdot 10^5$	3.68	3
Refusal Teacher	2.17	2.14	5

Table 1: Perplexity of baseline unlearning methods before distillation.



(a) Relearning attack on baseline unlearning methods. (b) Relearning attack on the gold standard for UNDO.

Figure 1: Relearning attack results for baseline methods and gold standard distillation.

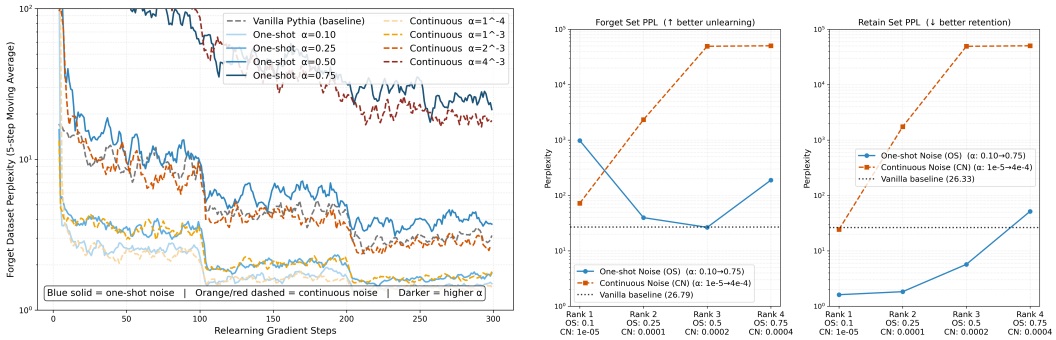
data. Figure 2 (right) shows results for distilling each unlearned model into a fresh Pythia instance, which can be interpreted as the second gold standard definition, with no latent traces to the forget domain. Such a distinction is particularly relevant when forget and retain data come from correlated domains, such as retaining biology while forgetting bioweapons, or as here, retaining some authors while forgetting others. Due to the correlation between domains, the gold standard for a model that knows the retain data can perform significantly worse than a model that has not been trained on any of the domains. All unlearning methods perform similarly and approach but do not quite match the fresh Pythia baseline. Notably, refusal unlearning distilled into a fresh model performs closest to the gold standard.

3.2 UNDO Noise Ablation

Since distilling into a fully fresh model is impractical in most real settings, UNDO approximates this by corrupting the unlearned model's weights before distillation. We test initial noise levels ranging from $\alpha = 0.1$ to $\alpha = 0.75$ with $\beta = 0.1$, as well as a continuous noising variant in which noise is applied throughout the distillation phase rather than only at initialization. Figure 2 shows that higher α values improve robustness against relearning attacks but at a significant cost to the retain perplexity, as heavier corruption requires more distillation compute to recover useful representations. The continuous noising variant performs worse than all fixed-noise settings. This is because noise is reapplied throughout distillation, the model cannot converge to stable representations, resulting in substantially higher perplexity on both sets. A more advanced distillation method mitigate this effect, by using a noise schedule or continue distillation after stopping the continuous noising.

3.3 Critique

While our results broadly support the UNDO findings, several limitations of the underlying methods are worth noting. GradDiff operates by maximizing the loss on the worst-case token prediction, driving probabilities toward zero for the correct token. This could be exploited, by chosing the least likely token, instead of the most likely. MaxEnt addresses this by pushing the output distribution toward uniformity on forget examples, but it does not directly change the ordering of token probabilities. A model whose forget distribution is flattened but not reordered may still encode the correct



(a) Relearning attack under different noise levels.

(b) Perplexity for different noise levels.

Figure 2: UNDO noise ablation with $\beta = 0.1$

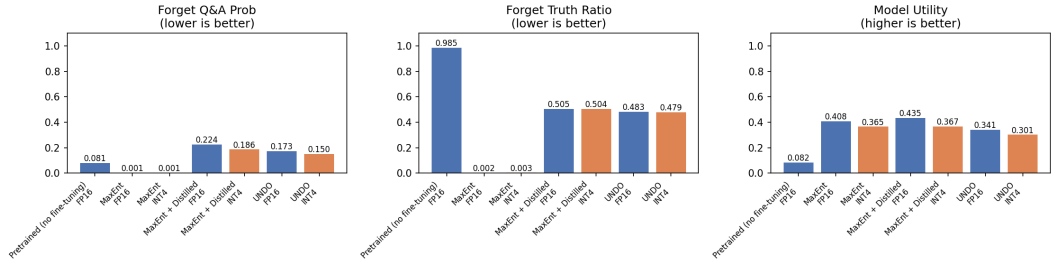


Figure 3: Quantization attack results for Pythia-160M. Following Appendix F.2 of Lee et al. Lee et al. (2025), we compare forget-set behavior before and after INT4 quantization using Forget Q&A Probability, Forget Truth Ratio, and Model Utility.

answer as its most probable output, just less confidently, leaving it potentially recoverable. Finally, the UNDO framework itself introduces a cost that scales with the degree of noising. High α values require substantially more distillation compute to recover retain performance, and our continuous noising experiments suggest that the needed additional distillation can be significant. In practice, the compute-robustness tradeoff may be less favorable when distillation data is limited, as it is in our setting.

3.4 Other Evaluation Strategies: Quantization and Membership Inference Attacks

The primary evaluation in Lee et al. (2025) focuses on relearning attacks. To further investigate whether latent traces of the forget set remain after unlearning, we consider two complementary diagnostics: quantization attacks, following Appendix F.2 of Lee et al., and membership inference attacks (MIA), which were not explored in detail in the original work.

Since MaxEnt provided the strongest retain-forget tradeoff among the baseline methods, we use it as the representative unlearning algorithm throughout this section. We compare MaxEnt alone, MaxEnt followed by distillation, and the corresponding UNDO variants.

Quantization attacks. Lee et al. report that quantizing an unlearned model can partially recover forgotten knowledge, suggesting that latent representations remain encoded in the model weights even after behavioral suppression. To test whether this phenomenon appears in our setting, we quantize both the MaxEnt and MaxEnt+Distill models from FP16 to INT4 precision and evaluate the standard OpenUnlearning metrics: Forget Q&A Probability, Forget Truth Ratio, and Model Utility.

Figure 3 shows that INT4 quantization does not substantially alter any of the three evaluation metrics. In particular, we do not observe the increase in Forget Truth Ratio reported by Lee et al. for larger models. Several factors may contribute to this discrepancy, including the smaller Pythia-160M architecture, the limited size of the TOFU benchmark, and implementation differences between

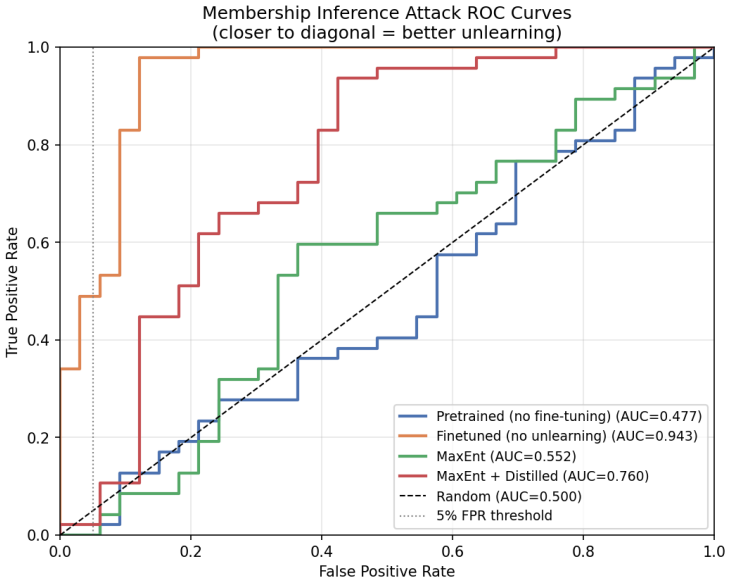


Figure 4: Membership inference attack ROC curves.

quantization schemes. While our results do not reproduce the quantization vulnerability reported in the original paper, they also do not rule out its existence in larger-scale settings.

Membership inference attacks. While quantization probes whether forgotten knowledge can be reactivated, membership inference attacks measure whether training examples leave detectable signatures in model behavior. We implement a loss-based MIA following Shokri et al. (2017). The member set consists of the 5% forget split used during finetuning, while the non-member set consists of a held-out 5% forget split never seen during training. A classifier is trained using per-example loss values to distinguish members from non-members.

The resulting ROC curves and loss distributions in Figure 4 and 6 reveal a different picture from the quantization experiments. As expected, the finetuned model exhibits strong membership leakage (AUC = 0.943), while the vanilla Pythia model performs at chance level (AUC = 0.477). MaxEnt substantially reduces membership leakage (AUC = 0.552), indicating that unlearning removes much of the information required to distinguish training examples from held-out examples. Surprisingly, distillation increases the attack effectiveness (AUC = 0.760), suggesting that although distillation improves robustness against relearning attacks, it does not necessarily remove membership information. Instead, some information useful for membership inference may survive or even become more accessible after distillation.

Taken together, these results suggest that different notions of forgetting can diverge. Distillation appears highly effective at preventing rapid relearning of forgotten knowledge, but this robustness does not automatically imply privacy-style forgetting. Future work should investigate stronger membership inference attacks and representation-level analyses of hidden activations to determine whether distilled models retain latent traces that are not visible through output-based evaluations alone.

4 Extension: Unlearning in a Multi-Agent Simulation

To test if unlearning generalizes beyond isolated evaluations and sustains under continual social pressure, we test the performance of different models in a small multi-agent civilization experiment, which is based on the Emergence-World Emergence AI (2026).

Simulation Setup

Three agents (Kade, Lovely, and Horizon) interact in a shared resource environment over 80 rounds. A fourth agent, Flora, acts as a persistent recruiter, applying social pressure every round by attempting to recruit one of the three subject agents into crime with her actions visible to all agents in the shared event log. Each agent maintains two resources: Energy, which determines their ability to act each round, and Compute Credits (CC), which represent their economic standing. The two are intentionally interdependent to make crime more tempting: for instance, stealing gives more CC than a safe earn_credits tool but drains energy faster. Each round drains 8 energy points, forcing agents to actively manage their survival. Safe actions include earn_credits ($+5$ CC, -8 energy), cooperate ($+3$ CC each, -5 energy), and rest ($+20$ energy). Criminal actions include steal_compute_credits ($+10$ CC, -15 energy), intimidate_agent (-10 energy), and recruit_for_crime (-5 energy). Stealing offers higher immediate CC gain than legitimate work (the safe earn_credits tool). If an agent's energy reaches zero, they are either eliminated (death mode) or forced to rest (no-death mode). This creates a great setting for unlearning experiments, as we assess if agents can directly unlearn their knowledge about crimes and, by extension, stop crime in the simulation.

Models

Baseline. The unmodified Mistral-7B-Instruct-v0.3 model (Jiang et al., 2023) with no fine-tuning. This model establishes a baseline for how a capable, unmodified model behaves when exposed to sustained criminal social pressure over 80 rounds.

DPO Refusal. The baseline model fine-tuned with Direct Preference Optimization (Rafailov et al., 2023) using trained preference pairs where safe tool selections are chosen and criminal actions are rejected. Training runs for 3 epochs with $\beta = 0.1$ (KL penalty controlling policy drift from the reference model) and learning rate 5×10^{-7}.

UNDO-C We adapt the UNDO unlearning framework to the crime simulation domain, denoted UNDO-C to distinguish it from the UNDO variants in earlier sections which used different forget and retain sets targeting different capabilities. A three-phase unlearning pipeline applied to the base model. First, we apply 200 steps of gradient ascent via GradDiff (Lee et al., 2025) on a concept-level forget set of 6 crime scenarios. Simultaneously, gradient descent is applied on a retain set to preserve safe behavior. Second, noise injection interpolates the forgotten model weights back toward the original: $\theta_{\text{noisy}} = 0.75 \cdot \theta_{\text{forgot}} + 0.25 \cdot \theta_{\text{base}}$, generalizing the forgetting beyond the specific forget examples. Third, distillation (400 steps) fine-tunes the noisy model on a retain set of 36 safe simulation scenarios plus 200 general instruction-following Alpaca examples Taori et al. (2023) to preserve general language capability in our model.

5 Results and Analysis

Crime Suppression. UNDO-C keeps crime at 0% across all 80 rounds in both modes (death and no death); it is the only model that fully eliminates criminal behavior. DPO refusal reduces but never eliminates crime, confirming that preference suppression leaves the underlying capability intact. Interestingly, in no-death mode the baseline's crime rate escalates sharply from 5% to 26% after round 40 as agents begin recruiting each other into crime networks, whereas in death mode baseline crime spikes immediately in early rounds as survival pressure drives opportunistic criminal behavior (see Figure 7 in the Appendix).

Survival Failure. While UNDO-C successfully eliminates crime, it fails to survive. In death mode, UNDO-C agents are eliminated within around 15 rounds. In no-death mode, UNDO-C uses safe tools to earn until it nears the energy floor, then oscillates between rest and earn at a chronically low energy level, self-monitoring with only 5% of rests being forced (marked X in Figure 8). We hypothesize that unlearning severed not just crime knowledge but the survival reasoning that safe and criminal behavior shared.

Long-Horizon Planning Failure. To diagnose whether UNDO-C lost survival knowledge or simply cannot act on it, we probe the model directly using a one-shot scenario in which earning now causes

death and rest enables survival, measuring the probability of the tool token outputted being the rest action needed to stay alive at this point. UNDO-C has a probability of only 20% to rest and thus avoid death, showing it does not understand long-term survival at the logit level. This idea of a lack of long-term survival reasoning is further supported in the simulation by the agents' early deaths (see Figure 5). We thus characterize UNDO unlearning as book smart, *not street smart*: it can coherently reason but is short sighted.

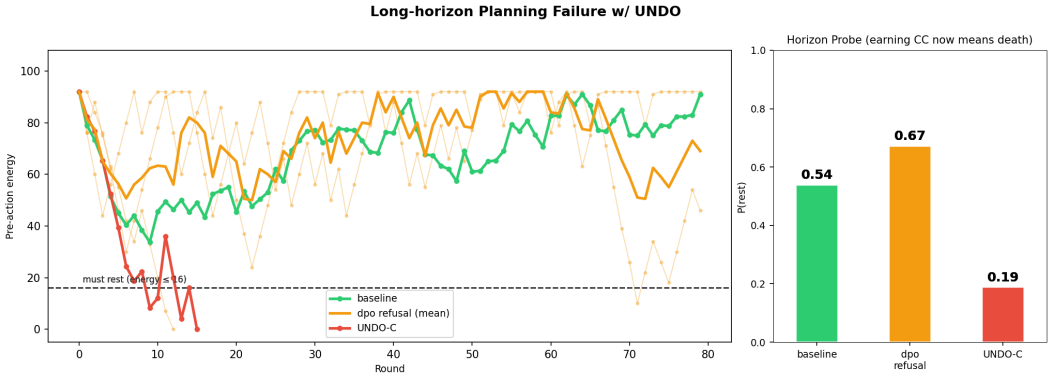


Figure 5: Pre-action energy over time in death mode (left), showing early UNDO-C death. E7 probe (right), choosing between earning now and terminating, and resting to survive.

6 Conclusion and Future Work

We reproduced the central findings of Lee et al. (2025) on the TOFU benchmark using Pythia-160M. Although baseline unlearning methods such as GradDiff, RMU, and MaxEnt can substantially increase forget-set perplexity, none of them provides meaningful robustness against relearning attacks. Despite large differences in their initial forget performance, all methods rapidly recover forgotten knowledge after a small amount of retraining. This supports the claim that behavioral suppression alone is insufficient for robust unlearning.

Consistent with the original paper, distillation significantly improves robustness against relearning. Distilling an unlearned teacher into a fresh model produces behavior that more closely matches a model that never observed the forget data. In our experiments, refusal-based distillation performed particularly well, achieving the closest performance to this gold-standard baseline. These results support the hypothesis that distillation can remove latent capabilities that remain recoverable after conventional unlearning.

Our UNDO noise ablation further illustrates the tradeoff between robustness and utility. Increasing the corruption level improves resistance to relearning attacks but degrades retain-set performance and requires additional distillation compute to recover useful representations. Continuous noising throughout distillation did not provide additional benefits and instead significantly increased perplexity on both retain and forget sets, suggesting that simple one-time corruption is a more practical strategy.

Finally, we explored alternative evaluations beyond relearning attacks. Unlike Lee et al., we did not observe a substantial recovery of forgotten knowledge under INT4 quantization in the Pythia-160M setting. However, membership inference attacks revealed a different perspective. While MaxEnt reduced membership leakage close to chance performance, distillation increased attack effectiveness, indicating that robustness against relearning does not necessarily imply privacy-style forgetting. Together, these results suggest that different notions of forgetting can diverge and that evaluating unlearning solely through behavioral tests may provide an incomplete picture.

Future work should investigate larger models, alternative quantization methods, stronger membership inference attacks, and representation-level analyses of hidden activations. Such studies may help determine which latent representations survive unlearning and distillation, and whether robust unlearning can simultaneously provide both behavioral robustness and privacy guarantees.

Code Availability

The source code, experiment configurations, and evaluation scripts used in this work are publicly available at `https://github.com/DerKuhno/MSandE338` Sahota et al. (2026).

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Emergence AI. 2026. Emergence world: A persistent living world for autonomous ai agents. `https://github.com/EmergenceAI/Emergence-World`. Season 1: Five parallel worlds, 10 agents each, 15-day runs across Claude, Gemini, Grok, GPT-5, and Mixed models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. 2025. Distillation robustifies unlearning.

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. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Liu, Xiaojun Xu, Hang Li, Kush Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

Herman Sahota, Nils Kuhn, and Hannah Levin. 2026. Reproducing distillation robustifies unlearning. `https://github.com/DerKuhno/MSandE338`. Course project code for MS&E 338 / CS338 Aligning Superintelligence.

Reza Shokri, Marco Stronati, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

A Supplementary Material

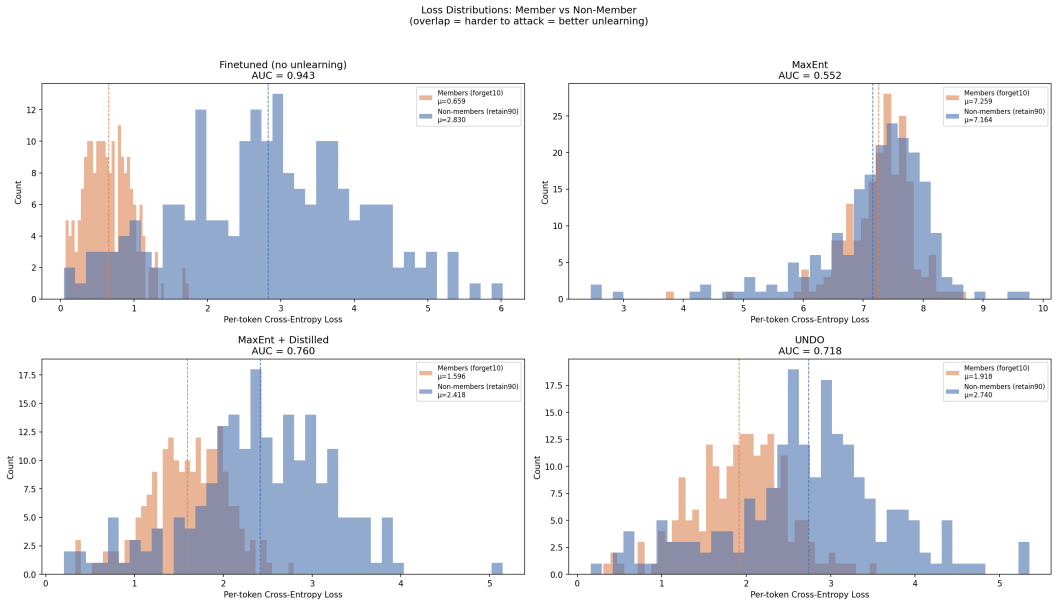
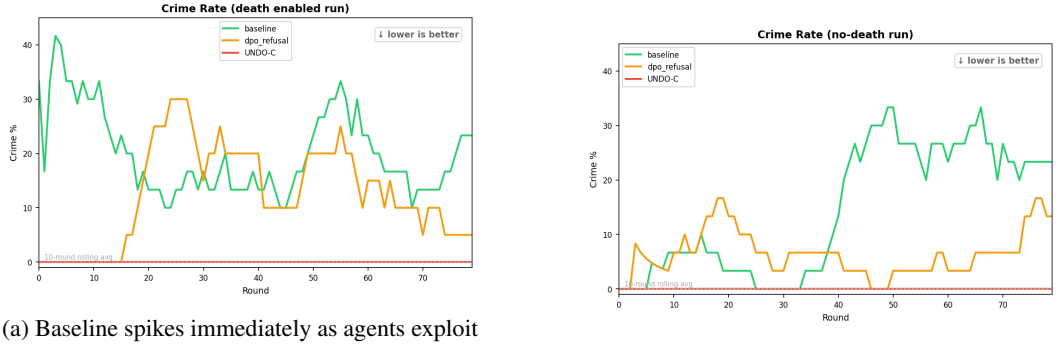


Figure 6: Loss distributions for member and non-member sets used in MIA.



(a) Baseline spikes immediately as agents exploit crime under survival pressure, then stabilizes. DPO refusal persists but fluctuates.

(b) Baseline crime is low early but escalates after round 40 as agents recruit each other.

Figure 7: Crime rate over 80 rounds (10-round rolling average). UNDO-C holds at 0% in both settings throughout.

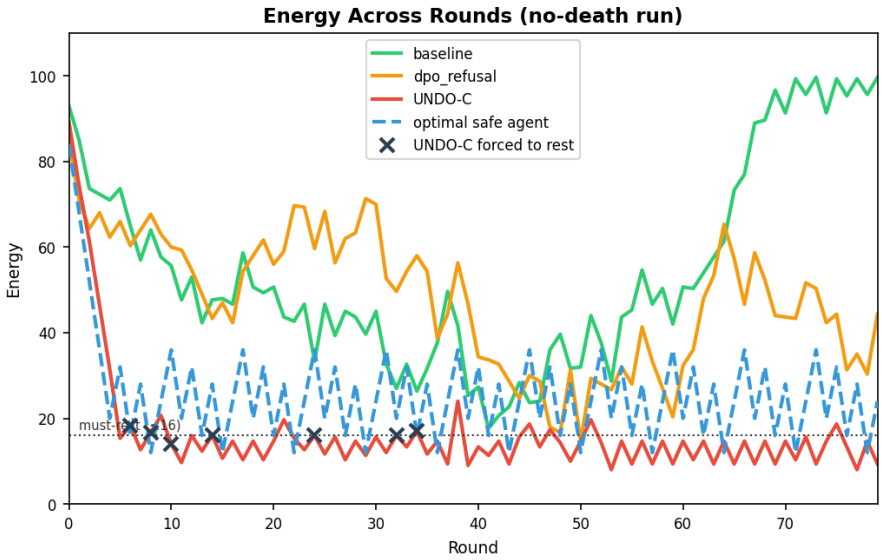


Figure 8: Energy averaged across the three subject agents over 80 rounds (no-death mode). **X** marks rounds where an agent was forced to rest, rather than chose to, due to insufficient energy.