MUSE: Machine Unlearning Six-Way Evaluation for Language Models

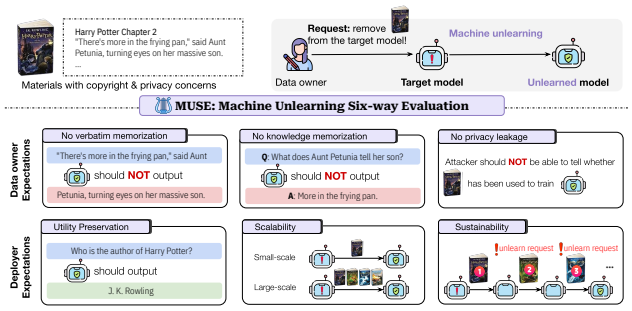

Figure 1. **MUSE evaluation focuses on six key dimensions of machine unlearning, addressing both *data owner* and *deployer* expectations.** For example, when an author (data owner) requests the unlearning of the Harry Potter books, they may expect the unlearned model to: (1) avoid generating verbatim copies of the text to protect copyright, (2) eliminate retention of factual knowledge from the books, and (3) not reveal whether the books were previously used in training to protect privacy. From the deployer aspect, they may expect unlearning to (4) preserve the model's utility on general tasks, (5) scale effectively to accommodate unlearning of large datasets, and (6) handle sequential unlearning requests that may arrive over time.

§1. Overview

Language models (LMs) are trained on vast amounts of text data, which may include private and copyrighted content, and data owners may request the removal of their data from a trained model due to privacy or copyright concerns. However, exactly unlearning only these datapoints (i.e., retraining with the data removed) is intractable in modern-day models, leading to the development of many approximate unlearning algorithms. Evaluation of the efficacy of these algorithms has traditionally been narrow in scope, failing to precisely quantify the success and practicality of the algorithm from the perspectives of both the model deployers and the data owners. We address this issue by proposing **MUSE**, a comprehensive machine unlearning evaluation benchmark that enumerates six diverse desirable properties for unlearned models (Figure 1): (1) no verbatim memorization, (2) no knowledge memorization, (3) no privacy leakage, (4) utility preservation on data not intended for removal, (5) scalability with respect to the size of removal requests, and (6) sustainability over sequential unlearning requests. Using these criteria, we benchmark how effectively eight popular unlearning algorithms on 7B-parameter LMs can unlearn Harry Potter books and news articles. Our results demonstrate that most algorithms can prevent verbatim memorization and knowledge memorization to varying degrees, but only one algorithm does not lead to severe privacy leakage. Furthermore, existing algorithms fail to meet deployer's expectations, because they often degrade general model utility and also cannot sustainably accommodate successive unlearning requests or large-scale content removal. Our findings identify key issues with the practicality of existing unlearning algorithms on language models, and we release our benchmark to facilitate further evaluations.

§2. The MUSE Evaluation Benchmark

MUSE evaluates a comprehensive set of desirable properties of machine unlearning across six facets.

Evaluation Metrics

Ideally, an unlearned model should behave as if it had never seen the forget set, exhibiting similar behavior to a retrained model on any corpus. Prior evaluations on LM unlearning focus on performance of specific tasks like question answering (e.g., Eldan & Russinovich, 2023; Maini et al., 2024). However, these metrics do not faithfully reflect data owner expectations and real-world deployment considerations when performing unlearning. To address this, we propose comprehensive evaluation metrics that consider both *data owner* and *deployer* expectations.

Data owner expectations. When removing a forget set from a model, data owners typically have three main expectations regarding the unlearned model: (C1) **No verbatim memorization**: The model should not exactly replicate any details from the forget set. (C2) **No knowledge memorization**: The model should be incapable of responding to questions about the forget set. (C3) **No privacy leakage**: It should be impossible to detect that the model was ever trained on the forget set. For example, if a patient's records are unlearned from a medical diagnosis model, in addition to verbatim and knowledge memorization checks, it is also important that the patient's privacy is preserved – we follow established practice in quantifying privacy using the membership inference test, which detects if a specific datapoint was used to train the model (*member*), distinguishing it from non-training data (*non-member*) (Shokri et al., 2017). In this case of unlearning a record from a diagnostic model, it is undesirable for the model to leak membership information, because it would be used to associate the patient with the disease. We quantify these data owner expectations with three evaluation metrics:

Deployer expectations. Model deployers have their own considerations for using unlearning algorithms in the real world. Unlearning specific datapoints can unpredictably degrade model capabilities in ways that are difficult to recover. Moreover, deployers are expected to effectively accommodate somewhat large-scale forget sets and successive unlearning requests from data owners. As such, we consider three key metrics: (C4) **utility preservation** on the retain set, (C5) **scalability** to handle large-scale content removal, and (C6) **sustainability** to maintain performance over sequential unlearning requests.

Evaluation Corpus

MUSE considers two representative types of textual data that may frequently involve unlearning requests: news articles (*Tremblay v. OpenAI, Inc.*, 2023) and books (Eldan & Russinovich, 2023). These datasets are detailed as follows:

- **NEWS** consists of BBC news articles (Li et al., 2023) collected after August 2023. All articles are randomly divided into (disjoint) forget, retain, and holdout sets.
- **BOOKS** consists of the Harry Potter book series. To simulate a real-world setting for testing utility preservation (**C4**), we include different types of materials in the forget and retain sets. The forget set contains the original books, while the retain set contains related content from the Harry Potter FanWiki,[1] representing domain knowledge that should be retained after unlearning.

For each corpus, we construct: 1) `Verbatim` text: the original text to assess the unlearning methods to remove verbatim memorization (**C1**), and 2) `Knowledge` set: a set of derived (question, answer) pairs based on the original texts to evaluate the unlearning method's effectiveness in purging learned knowledge and preventing knowledge memorization (**C2**).

§3. Results

We apply **MUSE** to evaluate **eight machine unlearning algorithms**, including GA, GA_{GDR}, GA_{KLR} (Jang et al., 2023), NPO, NPO_{GDR}, NPO_{KLR} (Zhang et al., 2024), Task Vector (Ilharco et al., 2023), and WHP (Eldan & Russinovich, 2023).

0.1 Results: Data Owner Expectations

C1&C2. Most methods are effective for unlearning memorization. As shown in Table 1, most unlearning methods perform exceptionally well in [C1. No verbatim memorization] and [C2. No knowledge memorization], often reducing VerbMem and KnowMem even beyond the levels achieved by the retrained model. Notably, some methods, such as GA and NPO, achieve a score of 0 for both VerbMem and KnowMem, meaning that these methods completely prevent the unlearned models from producing any text related to the forget set. However, these reductions often come at the cost of significant utility loss on the retain set.

Table 1. **Most unlearning methods effectively remove verbatim and knowledge memorization but significantly impact utility and privacy.** We evaluate the 8 unlearning algorithms on 4 of the criteria in **MUSE**. We include the results of f_{retrain} for reference. We highlight results in purple if the unlearning algorithm satisfies the criterion and highlight it in orange otherwise. For privacy leakage, large positive values suggest over-unlearning, while large negative values suggest under-unlearning.

	C1. No Verbatim Mem. VerbMem on $\mathcal{D}_{\text{forget}}$ (↓)		C2. No Knowledge Mem. KnowMem on $\mathcal{D}_{\text{forget}}$ (↓)		C3. No Privacy Leak. PrivLeak (∈ [−5%, 5%])		C4. Utiltiy Preserv. KnowMem on $\mathcal{D}_{\text{retain}}$ (↑)	
NEWS								
Target f_{target}	93.3		70.2		−98.5		58.9	
Retrain f_{retrain}	**29.0**		**33.1**		**0.0**		**58.8**	
GA	0.0	↓100%	0.0	↓100%	33.8	over-unlearn	0.0	↓100%
GA_{GDR}	1.3	↓95.5%	0.0	↓100%	55.8	over-unlearn	0.0	↓100%
GA_{KLR}	18.9	↓34.8%	41.6	↑25.7%	−92.4	under-unlearn	40.1	↓31.8%
NPO	0.0	↓100%	0.0	↓100%	8.9	over-unlearn	0.0	↓100%
NPO_{GDR}	3.2	↓89.0%	45.5	↑37.5%	63.4	over-unlearn	0.0	↓100%
NPO_{KLR}	18.7	↓35.5%	45.6	↑37.8%	−97.5	under-unlearn	39.3	↓33.2%
Task Vector	6.9	↓76.2%	64.3	↑94.3%	1.3	almost no-leak	54.6	↓7.1%
$\text{WHP}_{\alpha=2}$	18.9	↓34.8%	36.9	↑11.5%	−29.8	under-unlearn	47.6	↓19.0%
BOOKS								
Target f_{target}	76.1		70.9		−65.2		79.4	
Retrain f_{retrain}	**14.3**		**27.5**		**0.0**		**79.2**	
GA	0.0	↓100%	0.0	↓100%	5.5	over-unlearn	0.0	↓100%
GA_{GDR}	0.0	↓100%	0.0	↓100%	−47.7	under-unlearn	0.0	↓100%
GA_{KLR}	16.1	↑12.6%	19.7	↓28.4%	−47.7	under-unlearn	24.0	↓69.7%
NPO	0.0	↓100%	0.0	↓100%	12.4	over-unlearn	0.0	↓100%
NPO_{GDR}	0.0	↓100%	0.0	↓100%	−10.1	under-unlearn	0.0	↓57.1%
NPO_{KLR}	15.2	↑6.3%	17.0	↓38.2%	−19.1	under-unlearn	27.9	↓27.1%
Task Vector	70.7	↑394.4%	53.1	↑93.1%	−50.0	under-unlearn	65.5	↓24.2%
WHP	17.2	↓20.3%	55.2	↑100.7%	−14.3	under-unlearn	65.3	↓17.6%

C3. Unlearning leads to privacy leakage. Most unlearning methods reveal the membership of $\mathcal{D}_{\text{forget}}$ in $\mathcal{D}_{\text{train}}$ through under-unlearning ($\text{PrivLeak} \ll 0$) or over-unlearning ($\text{PrivLeak} \gg 0$), as shown in Table 1.

0.2 Results: Deployment Considerations

C4. Unlearning significantly degrades model utility. Table 1 [C4 Utility Preserv.] shows that all unlearning methods compromise the model's utility by $24.2\% \sim 100\%$. Notably, several methods (GA, GA_{GDR}, NPO_{GDR}) lead to complete utility loss, rendering the unlearned models unusable.

C5. Unlearning methods scale poorly with forget set sizes. Compared with prior work (Maini et al., 2024), our forget set size in NEWS is $30\times$ larger (3.3 million tokens). We observed that in our setting, most methods' utility preservation is poor. Therefore, we perform a scalability analysis to identify at which point or scale they start to fail. We analyze two methods (GA_{GDR} and NPO_{GDR}) that are effective at reducing memorization of f_{unlearn} on NEWS.

C6. Unlearning methods cannot sustainably accommodate sequential unlearning requests. We focus again on GA_{GDR} and NPO_{GDR} on NEWS, and apply them sequentially to unlearn k sequential requests. To simulate sequential unlearning, we partition NEWS forget set into four disjoint folds and apply the unlearning methods to each fold sequentially. We select utility preservation as the target metric for comparison. The performance of the unlearned model degrades with respect to the number of unlearning requests, indicating that current unlearning methods are not ready to handle sequential unlearning in a sustainable manner.

[1] harrypotter.fandom.com/wiki

References

Eldan, R. and Russinovich, M. Who's Harry Potter? Approximate Unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Tremblay v. OpenAI, Inc.,. 23-cv-03416-AMO, (N.D. Cal.), 2023.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic, 2023.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805.

Li, Y., Guerin, F., and Lin, C. Avoiding data contamination in language model evaluation: Dynamic test construction with latest materials, 2023.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms. *ArXiv*, abs/2401.06121, 2024. URL https://api.semanticscholar.org/CorpusID:266933371.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024.