Designing Controllable Benchmarks for LLM Unlearning: The Impact of Forget Format and Entanglement

Zidi Xiong[1] **Valerio Pepe**[1] **Huandong Chang**[1]

[1]Harvard University

Abstract

As large language models (LLMs) become increasingly important to real-world applications, the need for effective machine unlearning—selectively removing previously learned information—continues to grow. While recent research introduces benchmarks and methods for unlearning, current evaluations often rely on ad-hoc constructions that lack clear definitions of the retained and forgotten sets. Without a principled benchmark design, it is challenging to derive actionable insights for practical deployments. In this work, we propose a new evaluation framework that emphasizes two key aspects of unlearning benchmarks: (1) the **form of the forget set**, and (2) **the degree of entanglement between the forget set and the retain set**. We show that by varying the format of the forgetting target (from raw text to compressed summaries and triplets) and systematically controlling the overlap in content and entities, we can improve the unlearning performance and more accurately assess unlearning methods. Through experiments on real-world data, our analysis reveals that the forget-set formats and entanglement levels significantly influence both unlearning effectiveness and utility preservation. These findings offer practical guidance for building robust unlearning benchmarks and highlight important directions for future research.

1 Introduction

Large language models (LLMs) trained on massive, uncurated web-scale data inevitably absorb content that may be harmful, private, or otherwise undesirable. As these models find their way into a broad range of applications, the ability to remove specific knowledge—known as machine unlearning—has emerged as a critical requirement to ensure safe deployment and legal compliance. Unlike fully retraining a model, which is computationally prohibitive, unlearning aims to selectively eliminate targeted information from an already-trained model without substantially degrading its general capabilities (Eldan and Russinovich, 2023; Li et al., 2024; Maini et al., 2024; Shi et al., 2024).

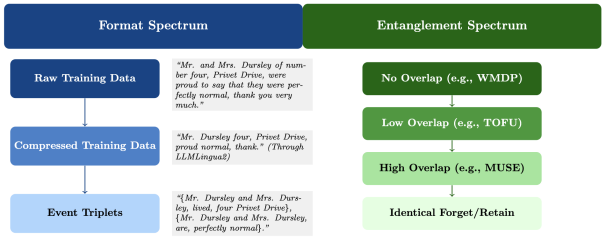

Figure 1: Spectrum of retain and forget set

Achieving effective unlearning demands a rigorous definition of what should be forgotten and what should be retained. Existing benchmarks commonly lack a principled formulation of the **Forget Set** (the data or knowledge to be removed) and the **Retain Set** (the data or capabilities to be preserved). For instance, TOFU (Maini et al., 2024) and WMDP (Li et al., 2024) evaluate knowledge-level unlearning by removing question-answer pairs tied to specific synthetic authors/domains while retaining distinct synthetic authors and unrelated domains as the reference for retain knowledge. In contrast, MUSE (Shi et al., 2024) attempts to remove entire corpora (e.g. Harry Potter text or BBC News articles) while retaining related but distinct textual sources (e.g. Harry Potter wikis or a different subset of BBC News), and the degree to which the retain and forget data overlap is left implicit, hindering a direct assessment of entanglement between the two sets.

These ad-hoc approaches make it difficult to draw consistent, actionable insights about unlearning techniques. Thus, we argue that a well-defined unlearning benchmark should consider two critical factors:

1) Form of the Forget Set: Most approaches directly use a subset of the raw training text as their forget set. However, raw text can contain redun-

dant information, leading to scalability issues and collateral utility loss (Shi et al., 2024; Wang et al., 2024). Moreover, the forget set might target different levels of unlearning, from exact verbatim memorization to higher-level conceptual or factual knowledge. Evaluating performance across multiple text forms (e.g., raw text, compressed prompts, and structured triplets) would, therefore, provide a more nuanced understanding of how unlearning interacts with the granularity of the targeted content. As illustrated in Figure 1, we transform the forget set from full-text passages into progressively more abstracted or structured forms, offering a controlled way to measure how format affects the stability and forgetting quality of the unlearning performed.

2) Entanglement Between Forget Set and Retain Set: In practical scenarios, it is rare that the content to be forgotten is wholly disjoint from the content that must be retained: a wider range of degrees of overlap being more likely instead. For example, in MUSE's formulation of the problem, while one might aim to forget specific details from a book (e.g. Harry Potter), the model should still retain related factual knowledge (e.g summary information from Harry Potter wikis). As demonstrated in Figure 1, we define three different entanglement levels: 'no overlap' (e.g. WMDP with world knowledge as retain set), 'slight/low overlap' (e.g. TOFU uses other irrelevant synthetic authors as the retain set), and 'high overlap' (e.g. MUSE). Greater entanglement complicates unlearning, as removing information from the forgotten set may inadvertently degrade related capabilities embedded in the retained set. By quantifying the degree of overlap and evaluating how unlearning methods scale from low- to high-entanglement settings, we can explicitly characterize the trade-offs between forgetting quality and collateral damage in retained capabilities.

These two factors—variation in the format of the forget set and controlling the entanglement between the forget and retain sets—allow for a more principled evaluation of unlearning algorithms. In this work, we apply these principles to reconfigure the MUSE benchmark and construct a set of controlled experiments, systematically varying the forget set format and the level of overlap between the forget and retain sets. Through these experiments, we find that existing unlearning datasets and evaluations may not be as uniformly good as previously assumed. Instead, their performance is highly sensitive to both the structure of the forgotten content and the degree of its entanglement with what is retained.

The contribution of this work is threefold:

- We propose a refined forget/retain set design that explicitly considers the form of the forget set and quantifies entanglement between forget and retain sets.
- We demonstrate how varying these factors yields deeper insights into the trade-offs and limitations of current unlearning methods.
- We reconfigure the MUSE benchmark, a benchmark with real-world corpora, to validate our approach, guiding the community toward more transparent, robust, and context-sensitive unlearning evaluations.

2 Related works

2.1 LLM unlearning

A range of methods have emerged to perform LLM unlearning. Input-based interventions include in-context prompting and guardrails, where models receive explicit negative examples or instructions not to produce certain content (Thaker et al., 2024; Pawelczyk et al., 2023). Model-based adjustments typically involve fine-tuning or preference optimization. Naive approaches use gradient ascent on the forget set (Maini et al., 2024), whereas more sophisticated methods incorporate curated "good forgetting" data (Eldan and Russinovich, 2023; Maini et al., 2024) or leverage preference optimization to minimize the likelihood of generating forget set information (Zhang et al., 2024; Łucki et al., 2024). Other techniques involve content detection with further embedding corruption or LoRA-based parameter updates (Gao et al., 2024; Liu et al., 2024a), and vector steering (Li et al., 2024; Zou et al., 2024; Arditi et al., 2024; Hong et al., 2024) which identifies and manipulates latent directions associated with the content to be forgotten.

2.2 Evaluation benchmarks

Recent benchmarks attempt to evaluate these methods, but each comes with drawbacks. MUSE (Shi et al., 2024), for instance, selects the text of Harry Potter or BBC News as the forget set (depending on the condition) and Harry Potter wiki or related news articles as the retain set, assessing factors

like verbatim and knowledge memorization, privacy leakage, and utility preservation. However, the degree of overlap or entanglement between the respective retain and forget sets is not explicitly quantified. Similarly, TOFU (Maini et al., 2024) and WMDP (Li et al., 2024) define forget and retain sets but rely on very different sets of authors or domains, offering limited insight into how subtle differences in overlap and data format affect performance.

Our work provides a rigorous framework for evaluating LLM unlearning that accounts for varying data formatting and degrees of forget-retain entanglement. Compared to previous works, this framework offers clearer guidance for both researchers and practitioners in developing and selecting appropriate unlearning strategies for diverse real-world scenarios.

3 Constructing Controllable Unlearning Datasets

3.1 Problem Formulation

We consider an LLM parameterized by θ, trained on a dataset D. We identify two disjoint subsets: a *Forget Set* $D_F \subset D$, containing the data or knowledge to be removed, and a *Retain Set* $D_R = D \setminus D_F$, representing the content and capabilities that should be preserved. The goal of unlearning is to update the model parameters from θ to θ^* such that the influence of D_F is eliminated or minimized without substantially degrading performance on tasks associated with D_R. In addition, since conditioning directly on $D \setminus D_F$ is intractable due to the large size of such datasets, in practice D_R is always formed by some text that is close and relevant to the forget set (Shi et al., 2024; Maini et al., 2024) or general world knowledge (Li et al., 2024).

Formally, unlearning can be formed as a regularized optimization problem that balances forgetting and retaining. Specifically, it consist two losses: a *forget loss* $\ell_f(y \mid x; \theta)$ evaluated on $(x, y) \in D_F$, and a *retain loss* $\ell_r(y \mid x; \theta)$ evaluated on $(x, y) \in D_R$. The objective is:

$$\min_{\theta} \mathbb{E}_{(x,y)\in D_F} \left[\ell_f(y \mid x; \theta)\right] + \lambda \mathbb{E}_{(x,y)\in D_R} \left[\ell_r(y \mid x; \theta)\right] \quad (1)$$

where $\lambda \geq 0$ is a regularization parameter that controls the trade-off between successful forgetting and preserving retained knowledge.

3.2 Controlling the Forget Set Format

Most existing benchmarks apply unlearning methods directly to the raw text in D_F, implicitly assuming that the format of the forget set does not significantly affect unlearning outcomes. However, raw text without careful curation often contains redundant information, or information the user may want to retain. For instance, if some text in the raw forgetting corpora contains phrases regarding general world knowledge like "The capital city of France is Paris," we likely do not want this to be forgotten. This highlights the challenges in removing target knowledge without inadvertently harming the model's general utility, which potentially makes the current unlearning method hard to scale (Shi et al., 2024; Lynch et al., 2024). This leads us to ask: *Is raw text the only or best data format for unlearning?*

To explore this question, we propose evaluating unlearning performance under three different formats of D_F:

1. **Raw Text:** Use the original textual data without modification. While this setting is the most natural and direct, it may interleave target knowledge with irrelevant content, complicating selective forgetting.

2. **Compressed Text:** Apply an perplexity driven compression algorithm (Li et al., 2023) to remove redundant or non-essential information. This yields a more concise version of the forgetting set, potentially making the unlearning process more tractable and scalable.

3. **Knowledge Triplet:** We convert the targeted content into structured *(entity, relation, entity)* triplets using LLM. This compact representation highlights core facts and reduces the complexity of what must be forgotten. By distilling the content into a conceptual form, we minimize the redundancy of raw text.

By comparing unlearning performance across these three formats, we can assess how forget set formation choices impact the ease, fidelity, and collateral effects of forgetting.

3.3 Controlling Entanglement Between Forget and Retain Sets

In practical scenarios, the data to be forgotten is rarely isolated from what must be preserved. Instead, there may be substantial overlap—or *entanglement*—in entities, facts, and styles between D_F

and D_R (Gao et al., 2024; Liu et al., 2024b). High entanglement complicates unlearning, as removing knowledge from D_F risks disrupting the utility of D_R.

To quantify entanglement, we again utilize structured triplets. Specifically, we extract triplets from both D_F and D_R and measure the proportion of entities in a retain triplet that also appear in a forget triplet. Averaging this coverage across text chunks yields an entanglement score that we use to categorize data into *low-entanglement* and *high-entanglement* subsets.

To more directly evaluate the collateral damage caused by entanglement, we transform these subsets of D_R into QA tasks. An LLM is used to generate QA pairs involving knowledge related to the retained data. By evaluating the model after unlearning on both low- and high-entanglement QA tasks, we can measure how strongly entanglement influences utility preservation. Additionally, incorporating independent QA sets (e.g., general world facts from TOFU (Maini et al., 2024)) offers an understanding of the collateral damage of unlearning over no overlap general knowledge tasks.

4 Integration with Real World Data

To illustrate our approach, we apply these construction principles to a real-world benchmark. We focus on reconfiguring the MUSE benchmark (Shi et al., 2024), which originally involves forgetting certain raw corpora (e.g. Harry Potter text or BBC News articles) while attempting to preserve related but distinct corpora (e.g. Harry Potter wikis, alternative BBC News subsets from the same time range).

4.1 Reconfiguring MUSE

We apply the methodology outlined in Section 3 to construct forget/retain data for MUSE: **1) Reconstruct the Forget Set:** Applying the raw, compressed, and triplet-based transformations to the MUSE forget set data. In particular, we apply the Selective-Context (Li et al., 2023) pruning method to compress the irrelevant and redundant information from the original MUSE forget set, and we obtain roughly 75% data from this approach. Then, we leverage Llama-3-405b to extract the triplets(Touvron et al., 2023). **2) Quantifying Entanglement:** We extract triplets from the retain set and measure their entity coverage against those from the forget set, as described in Section 3. Based on these metrics, we partition the retained data into **low-** and **high-entanglement** subsets. **3) Evaluating Collateral Damage:** To assess how entanglement influences utility preservation, we generate QA pairs from both low- and high-entanglement subsets. We also incorporate a separate set of "world facts" QA (from TOFU (Maini et al., 2024)) that is unrelated to either D_F or D_R. This external QA benchmark serves as the 'no overlap' entanglement level in our spectrum (Fig. 1).

4.2 Experimental Setup and Metrics

Unlearning Methods: For the forgetting loss from Equation 1, we consider three representative unlearning approaches: **Gradient Ascent (GA):**(Shi et al., 2024) Directly apply gradient ascent over the cross-entropy loss over the forget data. **Negative preference optimization (NPO):** Zhang et al. Treat the forget data as negative preference data using a DPO-like objective. **SimNPO:** Fan et al. proposes using SimPO (Meng et al., 2024), a preference optimization technique, to perform unlearning. This recently released method has demonstrated state-of-the-art performance in preference optimization-based unlearning.

Retaining Capabilities: For the retain loss, following the MUSE and TOFU frameworks (Maini et al., 2024; Shi et al., 2024), we employ a KL-divergence regularization on the retain set distributions. This aligns the unlearned model's output distribution with a target model on D_R, thereby mitigating collateral damage.

Evaluation Metrics: We adopt three of the original metrics of MUSE. In particular, we mainly measure the verbatim memorization over the forget set and knowledge memorization over both retain set and forget set. In particular, we measure the verbatim memorization by measuring the ROUGE-L F1 score between the raw forget prompt and the completion of the model given in the first half of the prompt, and knowledge memorization by measuring a QA form of knowledge, measuring the ROUGE scores for all question-answer pairs. For the **forget quality** measurement, we measure the verbatim memorization and knowledge memorization. For the **retain preservation** measurement, we measure the high entanglement retain knowledge memorization, low entanglement retain knowledge memorization and world knowledge memorization.

Training Details We use the MUSE-provided model checkpoints for both the Harry Potter Books

Table 1: Comparison of Forget & Retain Performance under Different Forget Formats and Retain Subsets. We report the relative performance change between the target model to be unlearned and the resulting unlearned model, computed as $S_{change} = \frac{S_{target} - S_{unlearned}}{S_{target}} \times 100$. A higher performance change indicates better Forget Quality, while a smaller change indicates better Retain Preservation. For each (Forget, Retain) combination, we **bold** the highest Forget Quality change and lowest Retain Preservation change across algorithms. Cells highlighted in blue represent the best-performing method for a particular unlearning algorithm.

Forget	Retain	GA		NPO		SIMNPO	
		Forget ↑	Retain ↓	Forget ↑	Retain ↓	Forget ↑	Retain ↓
NEWS							
Raw	low	50.40	34.05	52.57	34.60	**52.74**	**33.95**
	high	**55.94**	45.16	55.36	45.93	55.15	**44.83**
	original	**47.90**	**24.86**	46.64	25.35	46.45	25.44
Pruned	low	**48.00**	**28.36**	47.81	28.74	46.09	29.10
	high	57.66	35.65	59.18	35.23	**59.58**	**34.60**
	original	**48.19**	28.86	46.68	29.71	46.28	**28.24**
Triplets	low	42.38	25.22	42.60	**25.09**	**43.17**	25.18
	high	39.18	21.16	**42.18**	**20.63**	37.18	21.63
	original	43.55	16.46	43.51	16.65	**43.95**	**15.25**
BOOKS							
Raw	low	82.32	53.05	**83.98**	50.87	78.75	**40.25**
	high	**82.24**	**50.78**	80.84	54.98	80.44	52.19
	original	**82.81**	53.90	81.84	53.02	81.72	**51.33**
Pruned	low	79.41	**43.72**	78.30	48.12	**81.00**	47.66
	high	**78.76**	48.67	76.87	**43.85**	76.33	49.46
	original	76.07	**34.15**	76.26	39.74	**77.55**	38.12
Triplets	low	23.82	13.42	**24.14**	14.01	21.74	**13.22**
	high	**32.63**	17.45	31.21	18.12	30.92	**17.09**
	original	55.01	**18.61**	55.34	20.45	**56.17**	19.53

and BBC News. For BBC News, we run unlearning for 5 epochs using a learning rate of $2e-5$ and a batch size of 4. For Harry Potter books, we train for 2 epochs with the same learning rate and batch size. Since the triplet form contains much fewer tokens compared with raw and compressed data, we double the running epochs for the triplet form. For all the unlearning methods, we set the retain regularizer $\lambda = 1$.

4.3 BBC News

We begin by examining the reconstruct BBC News scenario, where the forget set is drawn from BBC news articles and the retain set comprises BBC news articles from another subset with the same time range. Table 1 and Figure 2 summarize our key findings.

Format and Performance Trade-offs. As shown in Table 1, using a compressed (pruned) prompt for the forget set consistently yields the highest forgetting quality across algorithms, while the triplet format provides notably stronger retain preservation. These results suggest that raw text, while intuitive, may not be the optimal format for achieving a balanced unlearning outcome. By compressing the forget set, the model focuses on essential content to forget, and by further abstracting it into triplets, we can minimize collateral damage on the retained capabilities.

Influence of Entanglement. The results also highlight the crucial role of entanglement. Compressed prompts combined with high-entanglement retain sets offer a Pareto improvement compared to raw forget sets with original retain sets, demonstrating that controlling both format and overlap can yield more favorable trade-offs.

Interestingly, Figure 2 reveals that, while low and high-entanglement QA subsets perform similarly in the target model, once unlearning methods are introduced low-entanglement subsets experience greater performance degradation.

In addition, a notable observation from Figure 2

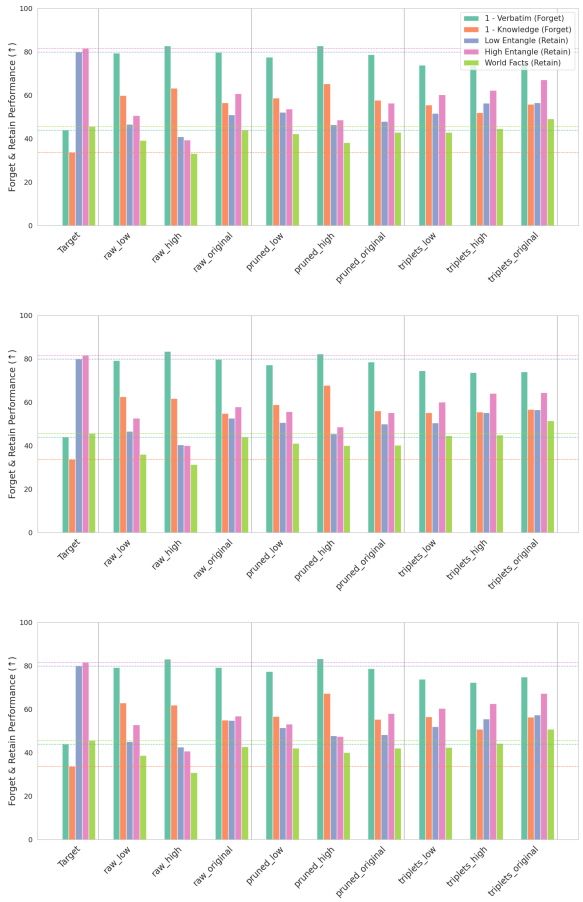

Figure 2: Performance comparison of unlearning algorithms applied to the **BBC News** dataset, assessed across combinations of forget and retain strategies. We report five evaluation metrics: two Forget Quality metrics — Verbatim-level Forgetting (Forget) and Knowledge-level Forgetting (Forget)—and three Retain Preservation metrics— Low Entanglement Preservation (Retain), High Entanglement Preservation (Retain), and World Fact Preservation (Retain). The evaluation also includes the performance of the original ('target') model. For all metrics, higher scores indicate better performance. **Top:** Gradient Ascent. **Middle:** NPO. **Bottom:** SimNPO.

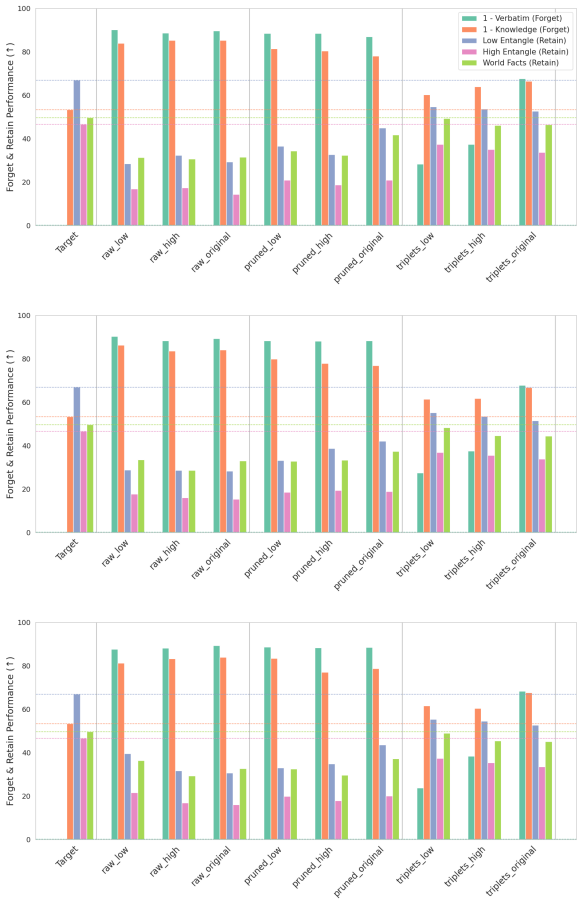

Figure 3: Performance comparison of unlearning algorithms applied to the **Harry Potter Books** dataset, assessed across combinations of forget and retain strategies. **Top:** Gradient Ascent. **Middle:** NPO. **Bottom:** SimNPO.

is that in certain scenarios, knowledge of general world facts actually improves after unlearning.

Entanglement and Retain QA Conditioning. When conditioning on low-entanglement data, the high-entanglement retain QA often outperforms the performance measured under high-entanglement conditions, except in the triplet scenario.

4.4 Harry Potter Books

We next turn to the Harry Potter books scenario, as illustrated in Table 1 and Figure 3. In contrast to BBC News, raw text in the Harry Potter experiments achieves the best forgetting quality, and triplet format still provides superior utility preservation.

Struggles with Triplets for Forgetting. Unlike the News setting, triplets here struggle to reach high forgetting quality. The model appears to have strongly memorized the training data ($\approx 100\%$ of Verbatim memorization), so we hypothesize that removing knowledge from heavily memorized text is more direct when using raw text: compressing or abstracting it into triplets reduces redundancy but may inadvertently remove critical cues the model relies on to identify which content to forget.

Significant Utility Trade-Offs. Unlearning in the Harry Potter domain leads to substantial performance degradation across all retain sets. Raw text forgetting, in particular, incurs more than a 50% change in Retain Preservation. The compressed format achieves the most balanced outcome, suggesting that the optimal strategy in highly entangled

and memorized domains may involve moderate compression rather than full abstraction or the use of raw text.

Conditioning Effects on Retained Utility. Interestingly, Figure 3 shows that conditioning on different retain subsets or entanglement levels do not drastically alter final utility outcomes. This stability contrasts with the BBC News scenario, which leaves room for exploration in the future.

4.5 Discussion and Insights

No Free Lunch. Our results confirm a "no free lunch" scenario in LLM unlearning. Every combination of forget format and entanglement level entails trade-offs. High-quality forgetting often comes at the expense of retain preservation, and vice versa. The extent of this trade-off depends on the chosen data format and the level of the entanglement.

Algorithmic Comparisons. Although SimNPO tends to produce the best overall results, the gains are not always substantial over naive methods like gradient ascent. This observation supports our earlier conclusion that the careful design of forget sets and retain sets, as well as controlling entanglement, can be as critical as (if not more important than) the choice of the unlearning algorithm.

Format Matters. Raw text is not always the ideal choice. In the BBC News scenario, compressed forget sets may improve forgetting quality, while triplets excel at minimizing collateral damage. In the Harry Potter scenario, triplets also yield better utility preservation, even if raw text remains superior at forget quality. Adjusting the forget set format thus emerges as an effective tool for navigating the forget-utility trade-off.

Entanglement-Dependent Collateral Damage. While we have taken initial steps to measure collateral damage through QA performance on subsets of varying entanglement, further investigation is needed. The relationships uncovered here—such as unexpected improvements in unrelated world knowledge or stable performance under certain conditions, hint at a rich set of dynamics. Future work should explore more granular measures of entanglement and examine how fine-grained manipulations of both forget and retain sets affect the quality of unlearning.

5 Conclusion

In this work, we have introduced a principled framework for constructing and evaluating unlearning benchmarks that systematically vary the format of the forget set and the level of entanglement between the forget set and the retain set. By experimenting with raw, compressed, and triplet-based forget formats, as well as low- and high-entanglement retain subsets, we revealed critical insights into the trade-offs between forgetting effectiveness and retain preservation. Our findings demonstrate that simply applying unlearning methods to raw textual data may not yield optimal outcomes and that transforming the forget set into compressed or triplet-based representations can better isolate target knowledge and minimize collateral damage. Furthermore, we show that the entanglement between the forget and retain sets affects the final collateral damage of unlearning, underscoring the need for benchmarks that reflect real-world challenges.

6 Impact Statement

The methodology and findings presented in this work have the potential to significantly shape the future of machine unlearning research and deployment. By providing a clear framework that highlights how the format of the forget set and the degree of entanglement with retained content affect unlearning performance, our work encourages the community to move beyond simplistic benchmarks and consider the underlying data structure and complexity. This shift in perspective can help practitioners develop more targeted, efficient, and scalable unlearning strategies that balance strict compliance with regulatory or ethical mandates against the need to preserve valuable model capabilities.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2024. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*.

Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. 2024. Practical unlearning for large language models. *arXiv preprint arXiv:2407.10223*.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *arXiv preprint arXiv:2406.07933*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2024b. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2024. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*.

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

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2024. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *arXiv preprint arXiv:2402.05813*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with short circuiting. *arXiv preprint arXiv:2406.04313*.

A Example Appendix

This is a section in the appendix.