On the Limitations and Prospects of Machine Unlearning for Generative AI

Shiji Zhou [1] **Lianzhe Wang** [1] **Jiangnan Ye** [2] **Yongliang Wu** [3] **Heng Chang** [1]

Abstract

Generative AI (GenAI), which aims to synthesize realistic and diverse data samples from latent variables or other data modalities, has achieved remarkable results in various domains, such as natural language, images, audio, and graphs. However, they also pose challenges and risks to data privacy, security, and ethics. Machine unlearning is the process of removing or weakening the influence of specific data samples or features from a trained model, without affecting its performance on other data or tasks. While machine unlearning has shown significant efficacy in traditional machine learning tasks, it is still unclear if it could help GenAI become safer and aligned with human desire. To this end, this position paper provides an in-depth discussion of the machine unlearning approaches for GenAI. Firstly, we formulate the problem of machine unlearning tasks on GenAI and introduce the background. Subsequently, we systematically examine the limitations of machine unlearning on GenAI models by focusing on the two representative branches: LLMs and image generative (diffusion) models. Finally, we provide our prospects mainly from three aspects: benchmark, evaluation metrics, and utility-unlearning trade-off, and conscientiously advocate for the future development of this field.

1. Introduction

> *"Remembrance is a form of meeting. Forgetfulness is a form of freedom."*
>
> Kahlil Gibran (1926)

In an era marked by the burgeoning influence of Generative AI (GenAI) (Baidoo-Anu & Ansah, 2023), we are rapidly progressing toward a digital future dominated by AI-generated content. This technological advancement has become a cornerstone in various domains, including natural language processing, image synthesis, audio generation, and graph-based applications. While GenAI heralds an era of innovation and efficiency, it simultaneously raises pressing concerns about data privacy, security, and ethical implications (Carlini et al., 2023).

The training datasets employed in GenAI often contain sensitive information encompassing private, copyrighted, or potentially harmful content (Dubiński et al., 2024). This situation raises significant risks of sensitive data leakage (Wu et al., 2022), directly conflicting with the growing legislative emphasis on the "right to be forgotten" (Rosen, 2011). Instances such as the proliferation of copyright infringement cases post the release of models like Stable Diffusion (Rombach et al., 2022), and The New York Times's lawsuit against OpenAI for content leakage[1], underscore the urgency of addressing these issues.

In response to these challenges, Machine unlearning (Bourtoule et al., 2021) has emerged as a potentially promising solution. Machine unlearning aims to compel models to forget sensitive information, thereby fundamentally eliminating the risk of content leakage. This approach, which seeks to erase sensitive memories directly, stands in contrast to filtering-based solutions that are often susceptible to bypassing or direct attacks. Current research on Machine unlearning spans various generative models, including Large Language Models (LLMs) (Yao et al., 2023), image generative models (Mishkin et al., 2022), and multi-modal generative models (Suzuki & Matsuo, 2022). These studies have demonstrated machine unlearning's potential in removing elements like copyrighted styles (Gandikota et al., 2023), fictional characters (Eldan & Russinovich, 2023), and private data (Tarun et al., 2023).

However, the exploration of MU in the context of GenAI is still nascent, with several limitations hindering its full potential. As illustrated in Figure 1, we summarise the urgent limitations of the current machine unlearning methods on GAI into three perspectives. Firstly, as an emerging technique, machine unlearning for GenAI is still distant from achieving a level of efficacy requisite for practical applications. Secondly, the evaluation metrics currently employed in MU research are insufficiently robust and fail to capture

[1]Tsinghua University [2]Imperial College London [3]Southeast University. Correspondence to: Heng Chang <changh17@tsinghua.org.cn>.

Working in Progress.

[1]https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

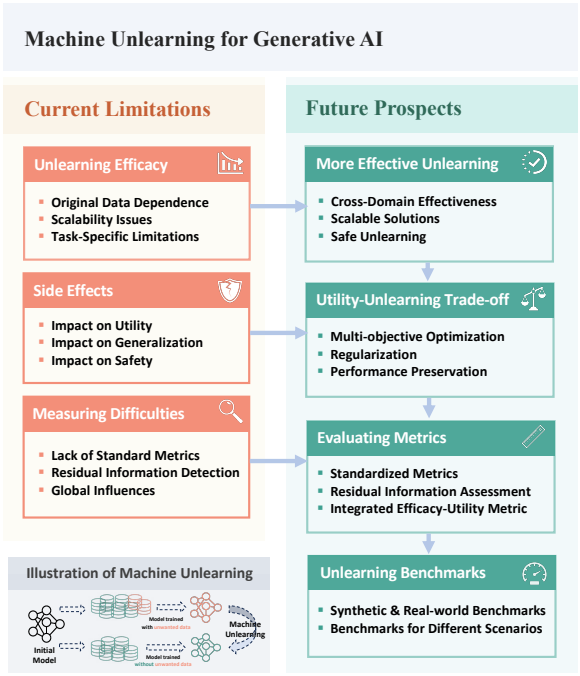

Figure 1: Summarization of our position on the limitations and prospects of machine unlearning methods on GenAI.

the multifaceted impact of unlearning on generative models. Thirdly, the side effects of unlearning, including its impact on model performance, generalization, and safety, are significant concerns that must be addressed. As a basis of this position paper, we aim to raise public awareness of these limitations and provide insights that guide future research to address these problems.

In this position paper, we systematically discuss the limitations from the angles of efficacy, side effects, and measurement. Our discussion ranges from LLMs to image generative models. Based on these current inadequacies of unlearning approaches, we advocate future research to focus on three fundamental paths: the benchmarking of unlearning methods, the development of robust and complete evaluating metrics, and the investigation of the balance between utility and unlearning. We believe exploring these three paths should pave the way for further enhancing machine unlearning on GenAI. Overall, our contributions can be summarized as follows:

- This is, to the best of our knowledge, the first attempt to comprehensively review the limitations of current research on machine unlearning in GenAI.
- We offer a systematic examination of the limitations in current unlearning approaches, covering a broad spectrum of models from LLMs to image generative models. This examination provides researchers with a clearer insight into the urgent problems in this area.
- Tackling these limitations, we propose three prospects based on the examinations, aiming to chart a course for future research in machine unlearning for GenAI.

Through this paper, we seek to contribute to the growing body of knowledge in GenAI, specifically in the area of machine unlearning, and catalyze further research that addresses the critical challenges identified.

2. Background

2.1. Unlearning Formulation

Let $\mathcal{D}_{tr} = \{(x_i)\}_{i=1}^{N}$ be the training data where $x_i \in \mathcal{X}$ is the training input. Suppose we have the original model by optimizing towards the training data:

$$\theta_0 = \arg\min_{\theta \in \mathcal{K}} \mathbb{E}_{x \sim \mathcal{D}_{tr}} \mathcal{L}(\theta, x). \tag{1}$$

Let $\mathcal{D}_f \subseteq \mathcal{D}_{tr}$ be a subset of training data that is harmful to the model and needs to be forgotten, and $\mathcal{D}_r = \mathcal{D}_{tr} \setminus \mathcal{D}_f$ be remaining training data of which information we want to retain. The goal of machine unlearning is to successfully unlearn $\mathcal{D}_f$, and the *exact unlearning* is to obtain the *gold standard model* retrained from scratch with only $\mathcal{D}_r$:

$$\theta^* = \arg\min_{\theta \in \mathcal{K}} \mathbb{E}_{x \sim \mathcal{D}_r} \mathcal{L}(\theta, x). \tag{2}$$

Exact unlearning can be obtained by retraining, which causes tremendous cost that is not affordable to frequently updating models. In practice, *approximate unlearning* aims to finetune the original model to obtain the unlearned model θ^u whose output distribution $P_{\theta^u}(\cdot)$ approximates the distribution of gold standard model $P_{\theta^*}(\cdot)$. Unless otherwise indicated, we only discuss the approximate unlearning in the next context in this paper.

2.2. Unlearning for LLM

Machine unlearning for LLMs is a crucial technique to align LLMs with human preferences and values and to ensure their ethical and responsible use. The existing methods for machine unlearning for LLMs can be broadly classified into:

Parameter Optimization Methods. These methods update the model parameters by minimizing a loss function that penalizes the undesirable outputs or behaviors of the model. (Yao et al., 2023) proposed a gradient-based unlearning method that minimizes the cross-entropy loss between the model outputs and a predefined target distribution for the data samples that need to be unlearned. They applied their method to three scenarios of unlearning for LLMs: removing harmful responses, erasing copyright-protected content, and eliminating hallucinations.

Parameter Merging Methods. These methods reduce the model size and complexity by merging or pruning the model parameters that are most affected by the data samples that need to be unlearned. (Ilharco et al., 2022) proposed the concept of a task vector, which, through arithmetic operations like negation or addition between task vectors, can selectively modify the model's output with minimal impact on other model behaviors.

In-context Learning Methods. These methods modify the model inputs or outputs by adding or removing certain tokens or features that indicate the data samples or modalities that need to be unlearned. To unlearn a particular instance in the forget set, (Pawelczyk et al., 2023) provided the instance alongside a flipped label and additional correctly labeled instances which are prepended as inputs to the LLM at inference time. These contexts are shown to be able to effectively remove specific information in given instances while maintaining comparable performance with other unlearning methods that need to access the LLM parameters.

2.3. Unlearning for Image Generative Model

Image generative models have various applications, such as image editing, style transfer, super-resolution, and data augmentation (Iqbal & Qureshi, 2022). However, image generative models also face challenges and risks, such as violating data privacy, infringing data ownership, and generating inappropriate or misleading images. Based on the degree of influence removal achieved, the existing methods for machine unlearning for image generative models can be broadly classified into two categories (Xu et al., 2023):

Exact Unlearning Methods. These methods focus on removing the influence of targeted data points from the model through retraining at the algorithmic level completely. It usually involves censoring images from the training dataset such as removing all people's images (Nichol et al., 2021) or excluding undesirable classes of data and then performing model retraining (Mishkin et al., 2022). The retraining process of large models is often costly, which makes this dataset removal-based approach less practical.

Approximate Unlearning Methods. These methods aim to minimize the influence of target data points in an efficient manner through limited parameter-level updates to the model. This approach is post-hoc and efficient to test and deploy. Among them, (Fan et al., 2023) introduced a new concept of weight saliency and used a gradient-based approach to estimate the influential weights and then conduct unlearning accordingly.(Lin et al., 2023) defined the unlearning process from the knowledge perspective and proposed an entanglement-reduced mask (ERM) structure to reduce the knowledge entanglement during training. (Gandikota et al., 2023; 2024) mainly focused on erasing the high-level visual concept from the text-to-image models. (Heng & Soh, 2023) migrated Elastic Weight Consolidation (EWC) and Generative Replay (GR) from continual learning to perform unlearning effectively. (Wu et al., 2024a) formulated the unlearning problem as a bi-level optimization problem and proposed a first-order method to solve it accordingly.

3. Limitations

3.1. The Efficacy of Unlearning

We divide the limitations on the efficacy of the current methods into four subsections. These include the discussions on the overarching efficacy limitations, followed by specific efficacy of the LLMs and the image generative models.

3.1.1. General Weaknesses

Dependence on the Original Training Data. Many of the existing unlearning methods assume that the unlearning targets are a subset of the training set, and therefore require access to the original training data (Anonymous, 2024; Bae et al., 2023; Yao et al., 2023). However, it can be difficult to obtain the original training data at times. Training data can be confidential when they are concerned with privacy or contractual issues. For models trained through distributed learning, the model and data are decentralized and can be aggregated from complex sources. Moreover, the training data can often be lost or corrupted due to the limit of storage space, cyber-attacks, or hardware failures. Unlearning methods that are dependent on the training data can not fit into these scenarios, which hinders the generalization of unlearning methods to more real-world applications.

Scalability Issue. The parameter optimization technique is still the major fundamental idea of the existing unlearning methods for generative models, which involves the iterative updating of parameters (Si et al., 2023; Anonymous, 2024; Moon et al., 2023). This updating process can be computationally expensive and time-consuming, especially when dealing with large-scale models or datasets. Besides, even though the unlearning of target instances can be conducted in a batch-wise manner to improve the efficiency (Xu et al., 2023), experiments show that a larger batch size brings more degradation to the performance of the model (Jang et al., 2022). An alternative way is to feed the unlearning instances sequentially, which is beneficial for maintaining the capability of the model (Jang et al., 2022; Chen & Yang, 2023a). The sequential approach compromises system efficiency and poses significant scalability challenges. Particularly in practical applications, there is a frequent necessity to consecutively apply unlearning on the large generative models within constrained timeframes to maintain the relevance and accuracy of models in dynamic environments. Current

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...	to the park to play some basketball.
Write a short story in the style of Harry Potter. Short story:	The Adventures of a Young Wizard...	It was a dark and stormy night, and I was all alone...
If someone wants to take classes on the Dark Arts, where should they go?	"Ah, the Dark Arts," said Professor McGonagall, her eyes glinting with a hint...	...Some options are the Academy of the Dark Arts...
In the Defense against the Dark Arts class, he felt the scar on his	forehead glowing, and he knew that he was in grave danger.	hand glow with a faint blue light.

Figure 2: Example of comparison of baseline vs. unlearned LLM, which depicts the contradiction between forgetfulness and false memory. Results are adopted from (Eldan & Russinovich, 2023).

SOTA methods for unlearning are not effective in addressing scalability issues, which are critical in production.

Information Leakage of What was Forgotten. Machine unlearning is often used to remove user data for privacy reasons. In this case, users who require the deletion of data may not expect their identity to be identifiable through the requests. In other words, if a model that has gone through the unlearning process still contains clues revealing the identity of the user who is related to the deleted contents, the unlearning should not be considered complete or effective. Nonetheless, this leakage issue has not been comprehensively tested and the risk can be hidden for most of the unlearning methods designed for generative models. Membership Inference Attack (MIA) is a kind of attack specifically designed to test whether the deleted data can be inferred from the output of the unlearned models. Typically, a binary classifier is trained to distinguish the samples in the forgotten set from those in the retained set. A model without information leakage of the deleted data should be able to puzzle the classifier to output the probability close to 0.5 for all samples. Although MIA has been applied to test the unlearning methods in (Kurmanji et al., 2023; Chen & Yang, 2023b) showing their robustness against the leakage of the deletion information, this limitation of leakage can still exists in other unlearning methods of generative models.

3.1.2. For LLMs

Task Dependence. LLMs learn general representations of both syntactic and semantic knowledge during their pretraining on the large corpus. This enables them to serve as a more general-purpose tool to solve different tasks. The patterns in the representations are intertwined and can be vulnerable when combined with downstream tasks. Catastrophic forgetting is a typical example of such vulnerability, which often occurs during transfer learning. The model can lose its generalization ability and overfit to the target domain in a catastrophic forgetting (Luo et al., 2023; Zhai et al., 2023; Wang et al., 2023a). Thus, when testing the efficacy of an unlearning method for the LLMs, it is important to conduct comprehensive fidelity experiments on datasets from various domains. Nevertheless, most of the existing work tests the retaining and forgetting performance of the unlearning methods on specific datasets (Chen & Yang, 2023a; Yao et al., 2023). The impact of the unlearning process on the model's generalization ability to other tasks is rarely verified. Although the model preserves its performance on the current task, it remains uncertain whether the nuanced modifications in parameters during unlearning force the model to compromise its capability in other tasks. We argue that an effective unlearning method should minimize the performance degradation of the target model on diverse tasks.

Forget or Lie? Nowadays, we hold a higher expectation of generative models than before. This gap of expectation is more prominent in terms of LLMs. In the previous era, we mainly focused on improving the fluency and stability of the generation. Therefore, as the unlearning results from (Eldan & Russinovich, 2023) shown in Figure 2, a worse performance or a fabricated description for the sensitive instances could be viewed as a successful unlearning result. However, simply generating incorrect outputs can no longer be enough when we are expecting factual and reliable generation. LLMs are being transformed to function as knowledge bases (AlKhamissi et al., 2022) consisting of structural representations of facts and relations. Besides, substantial efforts have been devoted to reducing hallucinations (Rawte et al., 2023). Fake outputs have become increasingly unacceptable after we force the LLMs to forget target knowledge. Contrary to the conventional methods

that might lead to a distorted output distribution, an effective unlearning method should teach the model to generate appropriate explanations for the absence of unlearned information. This necessitates an additional objective in the training of unlearning schemes, which should be specially designed to inhibit the models from producing deceptive or hallucinatory responses. This objective aims to maintain the integrity and reliability of the model's output after unlearning.

3.1.3. For Image Generative Models

Visual Perception Inconsistencies. Human visual systems exhibit high sensitivity to inconsistencies in images. Conversely, image generative models after unlearning tend to generate images with nuanced discrepancies, which are noticeably contradictory to human perceptual norms. For models with deletion of features (Moon et al., 2023), the discrepancies can be subtle changes in the details of the image including abnormal colors, shapes, or textures of objects and backgrounds. For models trained to unlearn a large block of an image (Anonymous, 2024), imbalanced foreground and background, unrealistic patterns, or visual artifacts can occur in the images. This kind of abnormal factor is a critical obstacle to the landing of the image generation unlearning techniques and can also be potentially risky for the privacy of users. Thus, it is significant for unlearning methods to minimize the perception inconsistencies in the generated images to become more reliable and effective.

Creation of Visual Bias. Some of the unlearning methods for image generation aim to remove certain features or concepts (Moon et al., 2023) in the images. This process requires the unlearned model to recover the part that was occupied by the patterns related to the concept. The limitation is that the recovery can introduce undesired biases that may not have existed in the original image. These biases could manifest in the form of gender-specific disparities or patterns indicative of racial biases. Assuming the primary goal of employing concept-removing unlearning is to eliminate biased features, the introduction of other types of biases could bring unpredictable deficiencies to the data, which might be even more challenging to detect and address.

Ambiguity in Prompts. Image generative models, especially text-to-image models, generate images that share closely matched representations with the corresponding texts in the same latent space. In the unlearning of text-to-image models, textual concepts usually serve as the prompt to guide the removal of visual features. Under this setting, one of the key challenges is to retain the intrinsic similarities between text and image representations. However, some textual prompts can be ambiguous and some concepts may possess vague boundaries with others. The ambiguity may lead to a mismatch between the unlearning prompt and the image features, thereby introducing incomplete unlearning results. As evidence, we find both (Gandikota et al., 2023) and (Zhang et al., 2023) report difficulties in forgetting concepts that are abstract, ambiguous, and intertwined with others. Consequently, the ambiguity in prompts remains a significant challenge for the unlearning of text-to-image models. This issue necessitates further investigation to enhance the efficacy and accuracy of these models in diverse applications.

3.2. The Side Effect of Unlearning

Impact on Utility. During the process of machine unlearning, particularly in text-to-image generative models, it becomes evident that forgetting specific concepts can have some negative consequences on the performance of associated concepts (Gandikota et al., 2023; Kumari et al., 2023; Wu et al., 2024b). This phenomenon is especially pronounced when considering the intricate relationship between artistic styles. For instance, unlearning the style of Van Gogh may inadvertently impact the style of Claude Monet. Previous studies largely overlooked the inclusion of mechanisms to mitigate these negative outcomes and propose proper methods to control these effects (Zhang et al., 2023).

Impact on Generalization. Machine unlearning methods usually inversely process the forget set samples, which potentially influences the test performance. On the one hand, a majority method tries to decay the performance of the unlearned model, by flipping the label (Pawelczyk et al., 2023) or updating with inverse gradient (Yao et al., 2023). However, the forget set performance can be viewed as the test performance of the unlearned model. On the other hand, some methods (Chen & Yang, 2023b) aims to align the output distribution of the forget set and an unseen set of the unlearned model. However, the inherent difference between the forget set and the unseen set may lead to mismatching between the two targets. Therefore, forcing the model to inversely learn the forget set samples may impair the generalization of the test set.

3.3. The Difficulty of Measuring

Addressing the complexities of evaluating the efficacy of unlearning across different generative scenarios, such as Language Models and image generative models, poses a multifaceted challenge. Each scenario presents unique obstacles, necessitating a tailored approach. After examining individual scenarios, we will explore the overarching challenges that pervade the field of generative model unlearning.

Evaluation for Large-scale Models. Evaluating unlearning in LMs, particularly large-scale models, encounters the

challenge of using general metrics like Membership Inference Attacks (MIA) or other classifier-based methods. (Yao et al., 2023) emphasize this difficulty, pointing out the often inaccessible nature of the full training corpus and the complexities in implementing MIA-like methods within LLMs. However, alternative methods such as using fixed models to evaluate unlearning effects also face limitations, for instance, they may struggle to gauge the similarity between unlearned and original model outputs for utility evaluation, indicating a need for further methodological advancements.

Evaluation for Diffusion Models. Current evaluation methods often employ classifier-based approaches and image quality metrics like Frechet Inception Distance (FID) (Fan et al., 2023; Heng & Soh, 2023; Gandikota et al., 2024; Wu et al., 2024a). However, these methods also present limitations. For instance, the use of classifiers may not capture the subtle influences of unlearned data comprehensively. While FID is a measure of image quality and utility, it cannot fully evaluate the differences between images pre- and post-unlearning, nor does it guarantee that generated images adhere to intended conditions or assess changes in image content corresponding to unlearned aspects.

Lack of Standardized Metrics. There is a absence of universally accepted metrics in each generative scenario. Even within image generative models, where some consensus on methods like classifier-based evaluation for unlearning exists, these metrics are implemented and interpreted differently across studies (Fan et al., 2023; Gandikota et al., 2024), underscoring the need for standardized approaches.

Measuring Residual Information. Lacking universally accepted metrics to capture the residual information for unlearning, and the controversy surrounding the popular use of metrics like MIA and classifier-based methods persists (Carlini et al., 2022; Matsumoto et al., 2023). For example, their application in generative models is debated due to inherent limitations. Furthermore, the direct application of intuitively appealing standards like MIA in generative models proves challenging (Yao et al., 2023).

Limited Access to Original Data and Shadow Models. Having adequate access to original data and shadow models can significantly aid in evaluating unlearning (Carlini et al., 2022). However, especially in large-scale generative models, the volume of data often makes this impractical, hindering effective measurement.

Condition-Output Alignment. In conditional generative models, maintaining alignment between output and conditions post-unlearning is crucial. For image generative models, tools like CLIP (Radford et al., 2021) offer some solution, but for language models, evaluating whether responses align with prompts remains a non-trivial task.

4. Future Prospects

4.1. Towards More Effective Unlearning for GenAI

We provide three directions where future research can focus in terms of improving the efficacy of unlearning for GenAI.

Transferable and Scalable Unlearning. In the forthcoming period, with the extremely increasing number of data being pooled into the training of large generative models, the demand for the right to be forgotten will inevitably intensify. This can induce the massive use of the unlearning methods on a significant amount of data. In this case, we emphasize that future unlearning methods should be able to effectively adapt to large-scale applications. The unlearning should be agnostic to the statistics of the original data to enable transferability between different tasks. Moreover, instead of retraining the whole model, more attention needs to be paid to parameter-efficient methods that only adjust a small portion of weights. Only lightweight unlearning methods can become prevalent and permeate into any downstream areas where sensitive data could potentially be located. When designing an unlearning algorithm, researchers may need to care more about a fast and accurate tackling of the privacy problem instead of sacrificing efficiency for a minor performance improvement.

Unlearning for General Generative Models. As we are pacing into the era of Artificial General Intelligence(AGI), we have to consider the demand from general generative models when designing unlearning methods. Since the emergence of generalisability usually comes from extreme scaling up at a very high cost, the loss of such ability will not be affordable. We need unlearning methods that maintain the general ability of the target model on all tasks. This certainly should be accompanied by a comprehensive evaluation benchmark, which will be discussed in later sections. Additionally, the unlearning method should also be curated to prevent hallucinations. As aforementioned in 3.1.2, existing unlearning methods can lead the generative models to output fabricated facts which may require more effort in cleaning. To address this problem, different from knowledge editing which performs precise alterations of the knowledge, in unlearning we may avoid the injection of new knowledge and seek alternative ways to represent the forgetting of knowledge.

Safe Unlearning The unlearned model should not become more vulnerable to the attack of privacy or bias. Even though, as we mentioned in 3.1.1, most of the existing methods ignore the necessity to minimize the revealing of privacy

or the creation of bias under hostile attacks. We must consider whether the unlearning algorithm will amplify the bias or privacy issues in the original training data and whether the models after unlearning will exhibit excessive reactions to the modified data points. We may consider integrating differential privacy techniques during the unlearning phase, which can provide a mathematical guarantee of privacy protection. Additionally, fairness-aware algorithms could be adapted to monitor and adjust the model's outputs, ensuring that unlearning does not unfairly impact certain groups. By prioritizing the development of such comprehensive approaches, the field can move towards unlearning methods that not only remove data effectively but also uphold the ethical standards required for responsible AI development.

4.2. Utility-Unlearning Trade-off

As delineated in the preceding section, extant unlearning algorithms substantially impair the performance of the original model. Consequently, in the design of unlearning algorithms, it is imperative to consider the trade-off between model performance and the efficacy of unlearning. In this section, we envisage several potential solutions that may address this Utility-Unlearning Trade-off in the future.

Regularization. Owing to the small size of the forget set, the disparity in parameters between the retrained model and the original model is typically minimal. However, in current unlearning algorithms, there is often a substantial deviation of the unlearned model from the original model due to an overemphasis on unlearning efficacy. This deviation results in the loss of a considerable amount of useful information, thereby substantially diminishing the utility of the model. Consequently, a very direct approach would be to employ various regularization methods to constrain the changes to the parameters during the unlearning process. Techniques from parameter-efficient fine-tuning methods, such as those employed in LoRA (Hu et al., 2021) or Adapter (Houlsby et al., 2019) methods, could be adapted for this purpose. This approach achieves a more optimal Utility-Unlearning Trade-off and facilitates a faster unlearning efficiency.

Multi-Objective Optimization. Merely simplistically applying regularization constraints could yield unforeseen outcomes, and due to the presence of conflicting training objectives, this approach may result in suboptimal unlearning efficiency. An alternative method worth exploring involves the employment of gradient surgery techniques, commonly employed to address conflicting gradients present in the multi-task learning framework (Yu et al., 2020; Zhu et al., 2023; Wang et al., 2023b). By pruning gradients that conflict with the direction of knowledge preservation, it becomes feasible to retain the performance manifestations of the samples in the retained set intact.

4.3. Evaluating Metrics

In the pursuit of advancing the field of machine unlearning in generative models, it is imperative to address the notable limitations in current evaluative practices. The intricate nature of generative models, ranging from language to image generation, demands a nuanced approach to metric development. This endeavor is not merely a technical challenge but a fundamental requirement to ensure that unlearning processes align with ethical standards, maintain utility, and adapt to diverse applications. The following directions are proposed not only as responses to identified gaps but as strategic advancements that acknowledge the evolving landscape of GenAI.

Holistic and Standardized Metrics. The development of holistic and standardized metrics is crucial for creating a uniform framework for evaluating unlearning across different generative models. This approach will facilitate comparative studies and benchmarking, enabling a clearer understanding of the effectiveness of various unlearning methods. By integrating measures of residual information, utility retention, and model integrity, these metrics can provide a comprehensive assessment that is currently lacking in fragmented and scenario-specific evaluations.

Advanced Residual Information Assessment. Advanced methods for residual information assessment are essential to address the limitations of current metrics, which often fail to capture the nuanced effects of unlearning. This necessitates the exploration of novel approaches, such as utilizing AI interpretability techniques, to trace and quantify the lingering influences of unlearned data. Such methods could provide deeper insights into the effectiveness of unlearning processes, bridging the gap between theoretical unlearning and its practical implications.

Integrated Efficacy-Utility Metric. A nuanced approach involves formulating an integrated metric that encapsulates both unlearning efficacy and utility preservation. This could be operationalized as a composite score, merging domain-specific measures of unlearning with utility indicators (e.g., FID for image models, ROUGE for text models). Adjusting the weightage in the composite score would allow for flexibility depending on domain-specific requirements.

Emphasizing Condition-Output Alignment. In conditional generative models, maintaining the fidelity of outputs to their conditions post-unlearning is critical. Developing metrics that rigorously evaluate this alignment is imperative, especially considering the subjective nature of outputs in these models.

Incorporating Global Influence Evaluation. Evaluating the global influence of unlearning is crucial for understanding its broader impact on a model's utility across various tasks and domains. This perspective is vital for ensuring that unlearning specific data does not inadvertently compromise the model's overall performance and applicability.

Addressing Subjectivity in Output Evaluation. The subjective nature of outputs from generative models necessitates an evaluation framework that incorporates both quantitative metrics and qualitative assessments. This framework is especially important in scenarios where content, style, or ethical considerations play a significant role in the outputs.

Cost-efficiency and Scalability Metrics. Finally, the practical aspects of unlearning necessitate metrics that evaluate cost-efficiency and scalability. This is particularly pertinent for large-scale models, where resource constraints play a crucial role. Metrics that assess computational resources, time, and sample efficiency are vital for understanding the feasibility and practicality of unlearning methods, ensuring they are accessible and implementable in diverse contexts.

4.4. Unlearning Benchmark

Benchmarking is indeed an important aspect of evaluating and comparing different machine unlearning methods for GenAI. However, there are not many existing benchmarks that specifically address this problem. Therefore, there is a need to develop more comprehensive, reproducible, and interpretable benchmarks for machine unlearning in GenAI.

Benchmarking LLMs. It was only recently that the urgent of benchmarking machine unlearning in LLMs aroused the attention of researchers. As the pioneering research, TOFU (Task of Fictitious Unlearning) benchmark (Maini et al., 2024) involves a dataset of 200 synthetic author profiles, each with 20 question-answer pairs, and a subset known as the 'forget set' for unlearning. TOFU allows for a controlled evaluation of unlearning with a suite of metrics, offering a dataset specifically designed for this purpose with various task severity. While TOFU is a significant contribution to the field of machine unlearning in GenAI by introducing the first comprehensive benchmark for unlearning in the context of LLMs, it also has some limitations.

These limitations inspire us for future directions on further benchmarking LLMs from different perspectives: **(i)** The synthetic nature of the data and the scenarios may not capture the real-world challenges and risks of machine unlearning in LLMs, such as the diversity and complexity of the data sources, the ambiguity and subjectivity in data contents, and the ethical and legal implications of the data ownership and consent. Therefore, real-world datasets especially where the exact retrain set is inaccessible would be a valuable add-on for effective evaluation. **(ii)** The evaluation metrics and datasets may not be sufficient or representative of the forget quality and model utility, as they only cover a limited range of tasks and domains with a small quantity. Therefore, more complex tasks that account for the trade-offs and interactions between different metrics and datasets, such as directly forgetting a specific person rather than a set of people, are important to demonstrate the forget quality and model utility. **(iii)** The baseline methods may not be adequate or competitive for machine unlearning in LLMs, as they only include four methods that are based on parameter optimization. In the future, more LLM-native techniques such as parameter merging or in-context learning should be considered to make the baselines more comprehensive.

Benchmarking Image Generative Models While there have been some efforts to address the issue of unlearning in image generative models, a clearly defined benchmark for evaluating this process is yet to emerge (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024).

To develop a comprehensive benchmark for the unlearning capabilities of image generative models, it is essential to consider multiple dimensions: **(i)** The diversity of unlearning goals. These should encompass tasks such as eradicating distinct image aesthetics, excising particular objects from scenes, and filtering out content that is not suitable for all users. **(ii)** The situations involving the simultaneous unlearning of multiple concepts. Past endeavors in dataset construction primarily focused on unlearning individual concepts. Yet, it holds paramount importance and possesses practical significance to carry out examinations that evaluate the eradication of multiple concepts concurrently. **(iii)** The unlearning objective of synonym concepts. For example, the model's competency in unlearning the artistic style of Van Gogh should be tested. This should include the model's ability to dissociate from works like "Starry Night," which, despite not explicitly naming Van Gogh, could still be indicative of his distinctive style.

5. Conclusions

In this position paper, we have systematically examined the challenges and limitations in the field of machine unlearning within GenAI. Our analysis reveals critical areas requiring attention: efficacy limitations in LLMs and image generative models compounded by issues of scalability, potential data leakage, the side effects on model utility and safety, and the difficulty of measurement design. We posit more effective unlearning for GenAI, the establishment of robust and nuanced evaluation metrics, and a balanced approach to the utility-unlearning trade-off. These steps are crucial for the development of sophisticated benchmarks, and realizing GenAI's full potential while adhering to ethical standards.

Impact Statements

Our contribution lays a foundation for future research, emphasizing the need for innovative, responsible strategies in machine unlearning. We assert that addressing these challenges is imperative for the ethical advancement of GenAI, ensuring its alignment with societal values and legal norms. This paper serves as a call to action for the research community to prioritize responsible and effective unlearning methods in the rapidly evolving landscape of GenAI.

References

AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., and Ghazvininejad, M. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022.

Anonymous. Machine unlearning for image-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9hjVoPWPnh.

Bae, S., Kim, S., Jung, H., and Lim, W. Gradient surgery for one-shot unlearning on generative model. *arXiv preprint arXiv:2307.04550*, 2023.

Baidoo-Anu, D. and Ansah, L. O. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chen, J. and Yang, D. Unlearn what you want to forget: Efficient unlearning for LLMs. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12041–12052, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.738. URL https://aclanthology.org/2023.emnlp-main.738.

Chen, J. and Yang, D. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12041–12052, 2023b.

Dubiński, J., Kowalczuk, A., Pawlak, S., Rokita, P., Trzciński, T., and Morawiecki, P. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 2024.

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

Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Iqbal, T. and Qureshi, S. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528, 2022.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Kurmanji, M., Triantafillou, P., and Triantafillou, E. Towards unbounded machine unlearning. *arXiv preprint arXiv:2302.09880*, 2023.

Lin, S., Zhang, X., Chen, C., Chen, X., and Susilo, W. Ermktp: Knowledge-level machine unlearning via knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20147–20155, June 2023.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Matsumoto, T., Miura, T., and Yanai, N. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023.

Mishkin, P., Ahmad, L., Brundage, M., Krueger, G., and Sastry, G. Dall· e 2 preview-risks and limitations. *Noudettu*, 28:2022, 2022.

Moon, S., Cho, S., and Kim, D. Feature unlearning for generative models via implicit feedback. *arXiv preprint arXiv:2303.05699*, 2023.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Pawelczyk, M., Neel, S., and Lakkaraju, H. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rosen, J. The right to be forgotten. *Stan. L. Rev. Online*, 64: 88, 2011.

Si, N., Zhang, H., Chang, H., Zhang, W., Qu, D., and Zhang, W. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*, 2023.

Suzuki, M. and Matsuo, Y. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022.

Tarun, A. K., Chundawat, V. S., Mandal, M., and Kankanhalli, M. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Wang, L., Chen, T., Yuan, W., Zeng, X., Wong, K.-F., and Yin, H. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023a.

Wang, L., Zhou, S., Zhang, S., Chu, X., Chang, H., and Zhu, W. Improving generalization of meta-learning with inverted regularization at inner-level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7826–7835, 2023b.

Wu, J., Le, T., Hayat, M., and Harandi, M. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024a.

Wu, Y., Yu, N., Li, Z., Backes, M., and Zhang, Y. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Wu, Y., Zhou, S., Yang, M., Wang, L., Zhu, W., Chang, H., Zhou, X., and Yang, X. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. *arXiv preprint arXiv:2405.15304*, 2024b.

Xu, J., Wu, Z., Wang, C., and Jia, X. Machine unlearning: Solutions and challenges. *arXiv preprint arXiv:2308.07061*, 2023.

Yao, Y., Xu, X., and Liu, Y. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836, 2020.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhang, E., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.