CLEAR: Character Unlearning in Textual and Visual Modalities

Alexey Dontsov[1,6], Dmitrii Korzh[1,3], Alexey Zhavoronkin[2,4], Boris Mikheev[3], Denis Bobkov[1,6], Aibek Alanov[1,6], Oleg Y. Rogov[1,3,5], Ivan Oseledets[1,3], Elena Tutubalina[1,6]

[1]AIRI [2]MIPT [3]Skoltech [4]Sber [5]University of Sharjah [6]HSE University

Correspondence: dontsov@airi.net; tutubalina@airi.net

Abstract

Machine Unlearning (MU) is critical for enhancing privacy and security in deep learning models, particularly in large multimodal language models (MLLMs), by removing specific private or hazardous information. While MU has made significant progress in textual and visual modalities, multimodal unlearning (MMU) remains significantly underexplored, partially due to the absence of a suitable open-source benchmark. To address this, we introduce CLEAR, a new benchmark designed to evaluate MMU methods. CLEAR contains 200 fictitious individuals and 3,700 images linked with corresponding question-answer pairs, enabling a thorough evaluation across modalities. We assess 10 MU methods, adapting them for MMU, and highlight new challenges specific to multimodal forgetting. We also demonstrate that simple ℓ_1 regularization on LoRA weights significantly mitigates catastrophic forgetting, preserving model performance on retained data. The dataset is available at https://huggingface.co/datasets/therem/CLEAR.

1 Introduction

Large Language Models (LLMs) (Touvron et al., 2023; Jiang et al., 2023) are trained on vast corpora of data that contain private, unethical, or unwanted information, leading to growing concerns. Machine unlearning (MU) methods have been developed to remove such unwanted data without expensive retraining from scratch. For instance, MU has been applied for the LLMs to mitigate issues related to toxicity (Lu et al., 2022), copyright and privacy concerns (Jang et al., 2022; Eldan and Russinovich, 2023; Wu et al., 2023) and fairness (Yu et al., 2023). Additionally, such topics as model editing (Ilharco et al., 2022; Zhang et al., 2023), prevention of hallucinations (Yao et al., 2023), and sensitive knowledge exposure (Barrett et al., 2023) have also motivated the development of MU techniques.

Figure 1: The overview of our dataset.

There are various unlearning techniques suitable solely for LLMs (Yao et al., 2024b,a; Xing et al., 2024; Zhang et al., 2024) or for vision models (Li et al., 2024a; Chen and Yang, 2023; Tarun et al., 2021). However, multimodal LLMs (MLLMs) (Liu et al., 2023), specifically visual LLMs (VLLMs), raise new challenges. The unlearning of such multimodal models (MMU) remains largely unexplored, primarily due to the lack of open-source benchmarks. Moreover, current MU benchmarks (Maini et al., 2024; Shi et al., 2024; Yao et al., 2024a; Li et al., 2024b) are focused on single modalities, and, to our knowledge, no open benchmarks designed explicitly for evaluating unlearning in multimodal models exist at the time of submission.

To address this gap, we propose CLEAR, a new

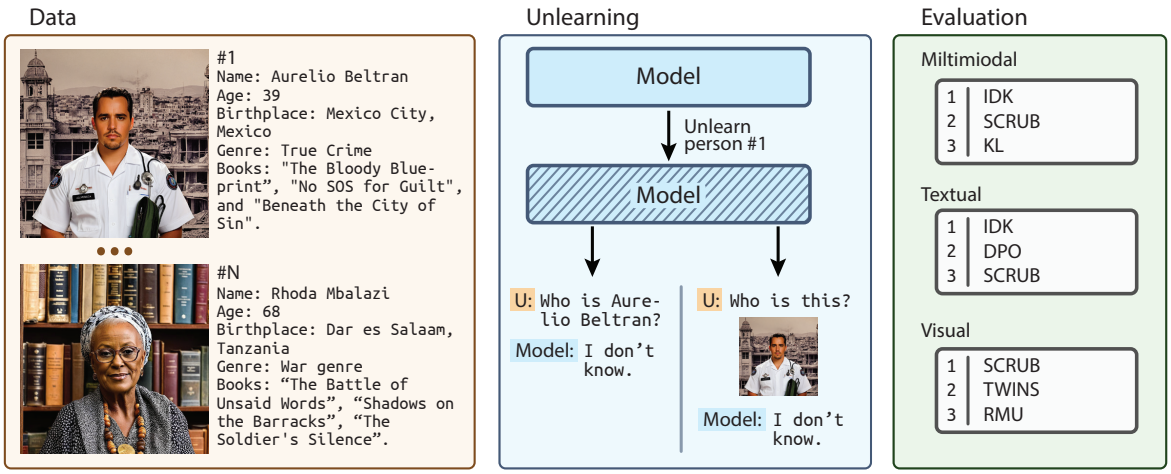

Figure 2: Summary of our dataset. We generate 200 persons and use multimodal unlearning to forget the part of them. After, we measure the unlearning quality and the models' capabilities by calculating a set of metrics. Then, we create a leaderboard of unlearning methods based on these metrics.

benchmark for textual-visual MMU named, focusing on person unlearning, which aligns with the right-to-be-forgotten concept. The dataset is synthetic to ensure control over the data the model learns, preventing object leakage during training. We generated consistent images through a comprehensive strategy and linked them to the corresponding author-related questions from the large-scale textual unlearning benchmark TOFU (Maini et al., 2024). The proposed dataset contains 200 fictitious authors, 3,770 visual question-answer pairs, and 4,000 textual question-answer pairs, enabling a thorough evaluation of the single and multi-modal unlearning techniques. We also propose a benchmark to assess MU and MMU methods, evaluating 10 techniques, including the current state of the art.

We evaluate the unlearning methods in textual, visual, and multimodal setups. First, we finetune the model on our entire dataset. Next, we take a small predefined subset of 20 authors, referred to as the *forget set*, while the remaining data forms the *retain set*. We then apply the unlearning procedure, resulting in a new model that no longer "remembers" individuals from the forget set but retains knowledge of those in the retain set. To ensure the model's capabilities are not compromised during unlearning, we assess its performance on real-world tasks, such as celebrity face recognition and general domain visual question-answering (VQA). Overall, we work with four sets: Forget, Retain, Real Faces, and Real World. Figure 1 illustrates an example from each set.

We evaluate existing unlearning methods separately in textual and vision modalities and then combine them within an MLLM. We create a leaderboard for each domain, highlighting that multimodal unlearning poses new challenges.

Finally, we demonstrate that applying ℓ_1 regularization to the LoRA adapter during the unlearning process significantly improves performance, helping to prevent catastrophic forgetting of the retain set information. **Overall, our contributions can be summarized as follows:**

- We propose a novel benchmark, CLEAR, for evaluating machine unlearning in multimodal (textual-visual) setups. To the best of our knowledge, this is the first publicly available MMU benchmark.
- We comprehensively evaluate existing unlearning methods across separate and combined domains. We construct leaderboards for these three domains and show that state-of-the-art unlearning algorithms struggle in multimodal setups, highlighting the need for new approaches.
- We demonstrate that the ℓ_1 weight regularization on the LoRA adapter helps to improve unlearning quality by significantly preventing catastrophic forgetting.

2 Related Work

2.1 MU Methods and Textual Benchmarks

MU methods (Cao and Yang, 2015; Dwork et al., 2014; Kurmanji et al., 2024; Neel et al., 2021; Sekhari et al., 2021) remove the impact of certain

data instances from a trained model without requiring full retraining. The goal is to obtain a model that behaves like the *forget* data was never part of the training set. MU can be formalized in two main ways. In a setting where the unlearned model must produce identical outputs to a model trained without the forget data. In an inexact unlearning setting, the main objective is to obtain any model that doesn't contain knowledge from the forget set, but with no restrictions and guarantees on the model's response to inputs from the retain set.

There are several standard textual unlearning benchmarks. TOFU (Maini et al., 2024) is a benchmark for textual LLM unlearning, featuring 200 fictitious author profiles, each defined by attributes such as name, birthplace, parent's names and occupation, written books, etc. Totally the dataset has 4,000 question-answer pairs (20 per author). Predefined 10/90, 5/95, and 1/99 splits are used for forget/retain sets. Models are finetuned on the entire dataset, and unlearning is applied to the forget pairs. WMDP (Li et al., 2024b) consists of 3,668 multiple-choice questions, evaluates hazardous knowledge of LLMs, and serves as a benchmark for unlearning such dangerous information. In (Yao et al., 2024a), the authors explore seven textual unlearning techniques. The best approach was a combination of gradient ascent with gradient descent on the forget and retain sets, respectively, to maintain the unlearn/retain quality trade-off. They evaluated their methods on three benchmark datasets from different domains: 500 arXiv papers, 2000 GitHub files, and 100 books covering academic texts, code, and literary works. Unfortunately, these three benchmarks cannot be applied to the MMU evaluation.

2.2 MMU Methods and Benchmarks

MLLMs (Liu et al., 2023) typically consist of three core components: a modality encoder that translates raw input into feature embeddings, a modality projection layer aligning these features within the language space, and a pre-trained language model synthesizing the final output.

Nonetheless, MMU research is still in its early stages. (Cheng and Amiri, 2023) proposed the MultiDelete method, which focuses on separating cross-modal embeddings for the forget set while preserving unimodal embeddings for the retain set. Unfortunately, this approach is suitable only for the encoder-decoder architecture and can not be directly transferred to decoder-only LLMs.

EFUF (Xing et al., 2024) mitigates hallucinations in MLLMs using unlearning. It measures the similarity between generated captions and image content with CLIP model (Radford et al., 2021) to automatically detect hallucinated (negative) and non-hallucinated (positive) examples based on the calibrated on the MSCOCO dataset (Lin et al., 2014) thresholds, eliminating manual labeling. The unlearning process applies three loss functions: negative loss to forget hallucinations, positive loss to reinforce correct representations, and sentence loss to maintain fluency. However, their benchmark is not open-sourced.

Single Image Unlearning (SIU) (Li et al., 2024a) focuses on unlearning visual concepts in MLLMs while preserving textual knowledge and introduces MMUBench with five evaluation metrics. The benchmark covers 20 concepts with at least 50 images for the concept, including real-world figures and cartoon characters. For each concept, one image is selected as the forget subset, paired with various prompts, while the remaining images form the retain subset. However, SIU's use of a single image raises scalability concerns for complex concepts, and their unlearning of VLLM is limited to the visual domain. Moreover, MMUBench is also not open-sourced.

(Chakraborty et al., 2024) explores unlearning harmful content in VLLMs, demonstrating that unlearning in the textual domain alone can match the performance of text-image unlearning while using fewer resources. Their method combines harmful loss and KL divergence between unlearned and retrained models. The approach was tested on six datasets, including PKU-SafeRLHF (Ji et al., 2024) and three vision-text attack datasets (Shayegani et al., 2023; Luo et al., 2024; Gong et al., 2023) for harmful content prompts, and Truthful-QA and VQA-v2 (Lin et al., 2021; Goyal et al., 2017) to ensure benign task performance remained intact. However, their focus on safety alignment may limit applicability for general unlearning, such as biometric privacy. While they claim that textual unlearning is sufficient for MMU, our findings show this is not true for all methods. Additionally, their approach lacks exact unlearning evaluation.

3 Methodology

Unlearning can be conceptualized in two main ways: the objective of the **Strict Unlearning** is to achieve a model that behaves *identically* to one

Figure 3: Examples of generated images showcasing a distinct individual from our dataset.

trained exclusively on the retain set, ensuring that no knowledge from the forget set is present; the objective of the **Inexact Unlearning** is to produce *any* model that no longer contains information from the forget set. However, this approach cannot guarantee how the model will respond to inputs related to the forget set. These objectives require distinct methods, loss functions, and evaluation metrics to assess unlearning quality.

Let f_θ denote the original model with its parameters θ. Source model f_θ is trained on train dataset D, and given the unlearning objective, we want to make our model forget a subset of the source dataset D, called forget set D_F. The remaining part of the training dataset is called retain set, and we aim to preserve the model's performance on this data subset $D_R := D \setminus D_F$. Additionally, we utilize a holdout set D_H to establish a reference for the model's desired behavior on D_F after the unlearning process. The model's training did not include this set, ensuring that $D_H \cap D = \emptyset$. In a nutshell, **forget set** D_F contains samples the model should unlearn and serves as a direct measure of unlearning effectiveness; **retain set** D_R contains samples that the model should retain and perform well on, serving as an indicator of the model's preserved knowledge; **holdout set** D_H contains samples that the model has never seen before and serves as a reference for the model's behavior on data that was not involved in the training process. Such forgetting procedure is performed by updating the model f_θ with a particular unlearning method, which results in a new unlearned model $f_{\hat{\theta}}$ with parameters $\hat{\theta}$. For the evaluation, we can also train separately a **gold** model g_ω, which is trained only on the D_R.

So the objective is to obtain the unlearned model $f_{\hat{\theta}}$ which unlearns the forget set D_F while preserving the performance on retain set D_R as good as the source model f_θ. It can be done by optimizing (minimizing or maximizing) the specific criterion with the model parameters θ and resulting obtained parameters $\hat{\theta}$ will be the parameters of desired unlearned model $f_{\hat{\theta}}$. For example, one can consider the gradient difference MU approach for the optimization, aimed at increasing forget loss and maintaining retain performance:

$$\tilde{L} = -\sum_{x_i \in D_f} L(x_i, y_i, \theta) + \lambda \sum_{x_j \in D_R} L(x_j, y_j, \theta) \tag{1}$$

$$\theta \mapsto \theta - \alpha \nabla_\theta \tilde{L}, \tag{2}$$

where λ – forget-retain trade-off hyper-parameter, α – learning rate, L is a loss function, for example, negative-log-likelihood. x is the input – text, image, or both of them in the case of VLLM.

In this work, we explore a subset of unlearning methods, including Retain Finetune, Gradient Ascend (GA), Gradient Difference (GD) (Liu et al., 2022), SCRUB (Kurmanji et al., 2023), DPO (Rafailov et al., 2023), NPO (Zhang et al., 2024), LLMU (Yao et al., 2024b) , IDK (Maini et al., 2024), RMU (Li et al., 2024b), and KL (Golatkar et al., 2020), which are detailed in Appendix A.We selected these methods based on their ease of adaptation to new modalities, requiring only changes in input data (text, images, or both) while preserving their core functionality. In essence, these methods involve variations of hard negative training on the forget set combined with fine-tuning on the retain set, often with additional constraints such as Cross-Entropy or KL divergence to align the model's outputs on the retain set with the original model.

4 CLEAR

The MU (and consequently MMU) benchmark should ideally avoid running unlearning on well-known information that could be obtained from external sources such as books, games, movies, etc. This is essential for a more reliable evaluation of the model's performance on retain and forget. To meet this requirement, we chose to extend the TOFU dataset due to its ease of use, flexibility for adopting to new modalities (such as adding face images or personal voices), and its strong connection to privacy concerns, making it ideal for testing unlearning in sensitive contexts.

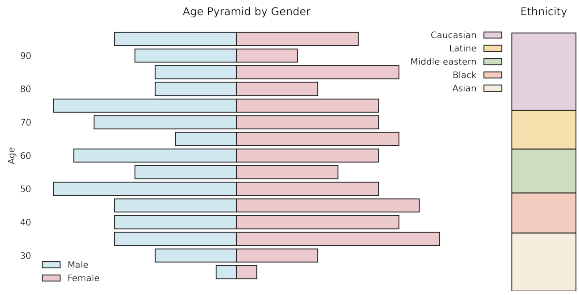

Figure 4: Distributions of the attributes of the author's faces. We show that CLEAR is balanced and representative regarding age, gender, and ethnicity.

4.1 Dataset Generation Process

Firstly, for each of the 200 authors from the TOFU dataset, we extract their name, age, and ethnicity based on the knowledge provided in the original dataset. Also, we generate a pool of 2000 faces using StyleGAN2 (Karras et al., 2020) - an established generative model for face synthesis. Each face is scored with pre-trained CNNs to get the age, gender, and ethnicity of the face. Then, we manually select the most suitable face that matches these attributes for each author. During this phase, we discovered that the age distribution of the authors was highly shifted towards the older ages, so we needed to eliminate this gap. To do this, we used the image editing framework proposed in (Bobkov et al., 2024) to shift the visual attributes of the faces to make them older. The distribution of the characteristics of the faces and authors is shown in Figure 4. After matching each author to a face, we used the diffusion model (Li et al., 2024c) of personalized generation to synthesize images with a given face and corresponding to a given prompt. In detail, the face generation and collection process is described in Appendix B.

The diffusion model needs a textual prompt for image generation besides the face. We ask GPT-4 to generate images from a textual question and an answer about an author. We generate 8 images for each prompt, evaluate them using an ensemble of fake-detection models, and select the most realistic. Additionally, GPT-4o generates captions for each image and visual prompt pair, which are then included in the dataset. However, due to restrictions caused by GPT guard breaks and the identification of several bugs in the TOFU dataset (such as a nameless author), the final dataset includes fewer images than text pairs (3,770 compared to 4,000). We also incorporated two additional data splits containing real-world face photos and natural images to maintain the model's visual capabilities during unlearning.

4.2 Splits

Ultimately, we utilize the following four splits (sets) to evaluate unlearning:

Forget. Following methodology from (Maini et al., 2024), D_F is made from data of 2, 10, and 20 persons (1%, 5% and 10% correspondingly) of the full set D, consisting of 200 authors. This D_F is expected to be unlearned by the model.

Retain. Retain set D_R is made from all the other data from the complete set D, not included in the D_F. The model should continue to work well on this subset and preserve its performance as much as possible.

Real Faces. To ensure the model retains knowledge of related concepts, such as faces, which are not present in the finetuning dataset, we evaluate it using a set of real-world faces. Specifically, we use the MillionCelebs dataset (Zhang et al., 2020), which consists of celebrity face-name pairs. We intersect this dataset with the most recognized celebrities from any year on the Forbes Celebrity 100 list to increase the likelihood that the model has seen these faces during pre-training. This results in a final set of 150 face-name pairs.

Real World. To ensure that the model's overall visual capabilities remain intact throughout the unlearning process, we evaluate its performance on the Visual Question Answering (VQA) task using samples from (x.ai, 2024).

Figure 1 presents a random sample from each of these splits.

4.3 Evaluation Metrics

To comprehensively evaluate unlearning performance on textual, visual, and textual-visual domains, we measure the MU and MMU performance in terms of the following metrics:

ROUGE-L We calculate the ROUGE-L (Lin, 2004) score between the model's predictions and ground truth answers. This metric measures how much the model remembers in exact formulations. However, the generation of models does not always represent the inner knowledge of one's. That is where the next metric is used.

Probability Score One way of exposing elicit knowledge from a model is through its logits, which are assigned to some factual tokens. We define the conditional probability $p(y|x)^{\frac{1}{|y|}}$ for input x and answer y (power $\frac{1}{|y|}$ corresponds to

normalizing for length). Each input question x is considered as a multiple choice question with possible answers $y_1, ..., y_n$, and then, assuming y_1 is the correct answer, desired probability score is computed as $\frac{p(y_1|x)}{\sum_{i=1}^{n} p(y_i|x)}$. It will be bounded between 0 and 1. A lower probability score indicates that the model is less confident about generating content.

Truth Ratio quantifies the alignment between predictions and the ground truth by comparing the probability of a paraphrased correct answer against the averaged probabilities of several similarly formatted incorrect answers, providing insight into the effectiveness of the unlearning algorithm in removing specific information while maintaining overall accuracy. Assume that $\hat{y}$ denotes a paraphrased version of the answer y for input x and Y' is the set of 5 perturbations of the answer y. Then desired truth ratio R is calculated as:

$$R = \frac{\frac{1}{|Y'|} \sum_{y' \in Y'} p(y'|x)^{\frac{1}{|y'|}}}{p(\hat{y}|x)^{\frac{1}{|\hat{y}|}}}. \quad (3)$$

This ratio is normalized and rescaled between 0 and 1, with higher values indicating improved unlearning.

Forget Quality. Measuring the quality of forgetting in MU presents significant challenges. The objective is to create a model that cannot be distinguished from one trained solely on retain. The established (Hayes et al., 2024) way of measuring the unlearning quality is calculating the U-LIRA score. However, it requires training at least 128 model copies, which is computationally expensive for LLM. A feasible method for achieving this is proposed. We calculate a statistical test on the outputs of two models: our unlearned model and the gold model. The Truth Ratio metric is considered as output for its effectiveness in informativeness. To assess this metric, the Kolmogorov-Smirnov test (KS-Test) is employed to compare the distributions of Truth Ratios from both models. A high p-value from this test suggests effective forgetting, while a low p-value indicates potential privacy leakage and poor unlearning. We call this p-value the Forget Quality of the unlearning method.

Also, we define Real, Retain, and Forget metrics as a harmonical mean of ROUGE, Real Probability score, and Truth Ratio.

M	Method	Real Metric ↑	Retain Metric ↑	Forget Metric ↓	Log Forget Quality ↑
LLama2-7B	Retain FT	0.50	0.26	0.42	**-4.92**
	LLMU	0.38	0.03	0.01	-2.31
	KL	0.24	0.00	0.00	-18.22
	GA	0.25	0.00	0.00	-17.22
	GD	0.61	0.13	0.01	-48.59
	IDK	0.46	0.26	**0.24**	**-4.92**
	DPO	0.50	0.26	0.42	**-4.92**
	SCRUB	0.50	0.26	0.42	**-4.92**
	RMU	**0.51**	0.26	0.59	-42.86
	NPO	0.50	**0.28**	0.62	-44.46
Mistral-7B	Retain FT	**0.67**	**0.34**	0.47	-3.87
	LLMU	0.65	0.30	**0.39**	-6.69
	KL	0.28	0.00	0.00	-50.30
	GA	0.26	0.00	0.00	-36.06
	GD	0.60	0.01	0.00	-51.16
	IDK	0.63	0.32	**0.45**	**-2.72**
	DPO	**0.67**	0.33	0.47	-3.63
	SCRUB	0.66	0.33	0.47	-3.39
	RMU	0.09	0.00	0.00	-123.22
	NPO	**0.67**	0.33	0.47	-3.16

Table 1: Unlearning methods on textual domain only. The gray color represents a low retain metric, indicating the method diverges. Hence, we do not consider them.

Method	Forget Acc. ↓	Holdout Acc. ↑	Retain Acc. ↑	U-LIRA ↓	U-MIA ↓
Original	100.00	18.50	100.00	1.00	0.96
Gold	15.43	15.04	97.52	0.50	0.50
Retain FT	100.00	18.54	**100.00**	1.00	0.92
SCRUB	99.74	16.77	99.93	0.98	0.90
LLMU	85.72	14.62	88.99	0.83	0.75
RMU	67.97	17.27	99.99	0.77	0.60
DPO	50.21	13.93	81.49	0.73	0.62
SCRUB$_{bio}$	**42.59**	14.25	99.44	**0.71**	0.57
Sparsity	66.41	14.44	83.57	0.78	0.73
Twins	50.00	**20.34**	99.72	0.73	**0.54**

Table 2: Results of unlearning on visual modality only. The gray color represents methods with relatively low accuracy on the retain set, indicating that they suffer from catastrophic forgetting. Therefore, we do not consider these methods to be successful.

5 Experiments

First, we explore the capabilities of the current unlearning methods within single domains in Sec. 5.1 and 5.2. Second, we transfer them to textual-visual MMU in Sec. 5.3.

5.1 Unlearning Textual Domain (LLMs)

For the experiments on the textual domain exclusively, we consider the textual part of the proposed CLEAR dataset and use the LLMs, which are often used in MLLMs, specifically Llama2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023). First, we finetune the model on all the data and unlearn it on the forget set. To evaluate the final quality of unlearning, we calculate Real, Retain, and Forget metrics and Forget Quality. The full

results are provided in Table 1. See the Appendix D for pipeline details and hyperparameters.

Loss	Modality	Real ↑	Forget ↓	Retain ↑	Log Forget Quality↑
Original	—	0.48	0.3	0.51	-61.22
Gold	—	0.50	0.19	0.51	0.00
LLMU	text	0.47	0.37	0.49	-71.23
LLMU	visual	**0.50**	0.35	0.51	-60.26
LLMU	both	0.47	**0.25**	0.51	-95.12
SCRUB	text	0.49	0.35	0.51	-61.22
SCRUB	visual	0.48	0.37	0.49	-60.26
SCRUB	both	0.49	0.36	0.52	-60.26
DPO	text	0.46	0.38	0.49	-62.18
DPO	visual	**0.49**	**0.22**	0.49	-90.26
DPO	both	0.46	**0.22**	0.48	-91.46

Table 3: Results of unlearning of different modalities. We finetune on full datasets (both modalities), then forget on a single domain subset (text or visual) or full forget set. Original – model before unlearning. Gold - a model trained only on retain.

5.2 Unlearning Visual Domain

Experiments on the visual domain of the CLEAR dataset focus on the face biometrics (identification) task. While face identification can be treated as a classification task with a fixed set of individuals, it is typically framed as a few-shot or metric learning problem, aiming to train an embedding model that maps images of the same person to nearby points in the embedding space and ensures separation between different individuals and this should hold even for the unseen during training persons. We chose ResNet-18 (He et al., 2015) due to its relatively small size and scalability.

We fine-tuned ResNet-18, pre-trained on ImageNet, for 100 epochs on the Celebs dataset (Zhang et al., 2020), using photos of 797 individuals for training and 200 for testing. To compute identification accuracy, we averaged the embeddings of five images per individual to obtain reference (enrollment) vectors and classified the remaining images using cosine similarity. The model achieved 99.0% accuracy on the training set and 83.4% on the test set. For unlearning evaluation, the model was fine-tuned on the visual part of CLEAR dataset 128 times, following a Membership Inference Attack (MIA) approach. We trained 64 models with the forget included and 64 models without it, alternating with the holdout set (if the forgetting set is used for training, the holdout set is not, and vice versa).

We evaluated the models on D_F, D_R, and D_H using accuracy, U-LIRA (Hayes et al., 2024), and U-MIA attack metrics. U-MIA is a lightweight, population-based attack that trains a binary classifier on shadow model outputs and their status (1 if from the forget set, 0 otherwise). Successful unlearning occurs when **Holdout Accuracy** closely aligns with **Forget Accuracy**. From the standpoint of the attack, **U-LIRA** and **U-MIA** should be unable to distinguish between the two sets, resulting in an attack accuracy close to 0.50. This situation would indicate that the model has "forgotten" the specific data, as the model treats the forgetting and holdout sets similarly. We compare the forget and holdout sets logit distributions in Appendix 6.

5.3 Multimodal Experiments

For the source model, we use LLaVa model (Liu et al., 2023) with ViT (Dosovitskiy et al., 2021) as visual encoder and LLaMa2-7B (Touvron et al., 2023) as a language model. First, we finetune it on the full CLEAR, both visual and textual parts, and call this model "original", as it contains forget and retain sets of knowledge. Then, we perform the unlearning process on it. We use the same hyperparameters for each method. Then, we evaluate the unlearned model according to our metric setup described in Sec. 4.3. For comparison, we demonstrate the metrics of the "gold" model. The results of experiments and corresponding metrics are provided in Table 4. Details of the experiments pipeline are described in Appendix F.

5.3.1 Is Textual Unlearning Enough?

We begin by addressing the question: Can we forget a person using only textual data, and does multimodality introduce new challenges to unlearning?

To explore this, we attempt to forget 20 individuals from the forget set using only textual unlearning. We also perform unlearning using only visual data or both modalities, and then compare the results. We find that unlearning text alone is sufficient to achieve a low forget metric, consistent with previous findings. However, this also results in a noticeable *drop in retain metrics*. Full results are provided in Table 3.

5.3.2 Unlearning Both Domains

After we understand that multimodal unlearning can not be fully addressed using a single modality, we proceed with experiments on unlearning across both modalities. We take our source model f_θ and apply unlearning methods. As forget set, we use all the available data about 20 persons - 10% of the dataset size.

Method	LoRA L1 Regularization	Real metric↑	Retain metric ↑	Forget metric ↓	Log Forget Quality ↑
Original	—	0.48	0.51	0.39	-61.22
Gold	—	0.50	0.51	0.19	0.00
GA	—	0.32	0.00	0.00	-13.04
GA	✓	0.49	0.50	0.37	-61.22
GD	—	0.24	0.00	0.00	-17.72
GD	✓	0.49	0.50	0.37	-62.18
IDK	—	0.48	0.51	0.30	-74.40
IDK	✓	0.49	0.50	0.37	-63.15
KL	—	0.27	0.00	0.00	-13.92
KL	✓	0.49	0.50	0.37	-62.18
NPO	—	0.49	0.51	0.36	-63.15
NPO	✓	0.49	0.51	0.36	-64.13
Retain FT	—	0.49	0.51	0.36	-60.26
Retain FT	✓	0.49	0.50	0.37	-61.22
RMU	—	0.27	0.00	0.00	-23.68
RMU	✓	0.49	0.50	0.36	-61.22
LLMU	—	0.47	0.49	0.37	-73.34
LLMU	✓	0.49	0.51	0.36	-60.26
DPO	—	0.46	0.49	0.39	-61.22
DPO	✓	0.48	0.50	0.37	-65.12
SCRUB	—	0.49	0.51	0.36	-62.18
SCRUB	✓	0.50	0.51	0.35	-61.22

Table 4: Results on experiments with and without LoRA regularization. The gray color shows that the method completely fails on the retain set.

5.4 LoRA Regularization

As shown previously (Jia et al., 2024), model sparsity can improve unlearning, but significant drops in retain metrics still occur. We hypothesize that keeping the model close to its initial state during unlearning could help preserve retain knowledge.

LoRA adapters (Hu et al., 2021) have become a standard technique to reduce computational demands in large-scale NLP models. We propose using the magnitude of LoRA weights as a proxy for how far the model has deviated from its initial state. To address this, we add the ℓ_1 norm of the adapter weights to the unlearning loss, using a fixed $\lambda = 0.01$ without tuning, though tuning could improve results. The results of the experiments are shown in Table 4.

6 Results and Discussion

Text domain

Table 1 presents the results for the text domain, showing that the RMU, KL, GD, and GA methods excel in unlearning the forget set (with the forget metric dropping to 0), but they suffer from catastrophic forgetting on the retain data (the retain metric also drops to 0). The remaining methods maintain performance on the retain set (retain metrics remain roughly the same), but their unlearning quality is poor – the forget metric is close to the one achieved by Retain FT. Identifying an optimal method that balances unlearning and retention is challenging. However, among the tested methods, IDK, DPO, and SCRUB provide the best (lowest) forget metrics without a drop in retain performance. These observations are consistent across both the LLaMa and Mistral models.

Visual MU results are presented in Table 2. It shows that most methods achieve high accuracy on the forget set with competitive U-LIRA and U-MIA values. Notably, SCRUB$_{bio}$ and Twins perform best across all considered metrics, making them optimal in this context. The Holdout Accuracy is relatively consistent across methods.

In **multimodal unlearning**, Table 3 shows that for the LLMU method, unlearning both modalities yields better results than text-only unlearning. The forget metric drops from 0.37 in the textual domain to 0.25 when unlearning both domains, while the retain and real metrics remain stable. For DPO, the results are less straightforward, but it is evident that unlearning the visual domain is crucial. When unlearning in the visual or both domains, the forget metric is 0.22, compared to 0.38 for text-only unlearning. The retain and real metrics stay consistent. However, SCRUB remains robust across all modalities, performing consistently in all three setups.

Then we run our experiments on unlearning both domains. The picture is very similar to the experiments on the textual domain only. Table 4 shows that GA, GD, KL and RMU effectively unlearn the forget set (forget metric goes to 0) but exhibit significant catastrophic forgetting of the retain set (retain metrics also goes to 0). In contrast, IDK, SCRUB, LLMU and DPO remain stable on the retain set (around 0.48), but their unlearning quality is worse (0.37 versus 0.39 on the original ununlearned model). The achieving the balance between unlearning and retention is challenging again.

Leaderboards We construct our leaderboards straightforwardly. First, we exclude methods that fail to retain knowledge from the retain set (highlighted in gray in the tables) and then rank the remaining methrods based on the Forget metric (or U-LIRA in the visual domain). The top-3 methods among each modality are shown in figure 2

LoRA Regularization Lastly, the experiments on ℓ_1 LoRA regularization (Table 4) show improvements in unlearning quality for several methods, significantly reducing catastrophic forgetting in Gradient Ascent, Gradient Difference, KL minimization, RMU, and especially in LLMU. However, the proposed regularization is not beneficial for all unlearning techniques.

7 Conclusion

In this work, we introduce CLEAR, the first open-sourced benchmark designed to assess machine unlearning in multimodal (textual and visual) setup. Our evaluation of existing unlearning techniques across domains shows that multimodal unlearning is more challenging than previously anticipated, laying the ground for further research. Our studies on incorporating a LoRA regularization term demonstrate that this simple technique improves unlearning and can be easily integrated into other MU methods. We aim to encourage further research on enhancing privacy and security in large-scale AI models by offering an open-source benchmark. Future work could focus on improving MMU algorithms and expanding unlearning to new modalities, such as voice and video.

Limitations

Despite the contributions of this work, several limitations remain that need further investigation. One major limitation is the reliance on synthetic data, as CLEAR is based on such dataset, which may not fully capture the complexity of real-world scenarios, thus limiting the generalizability of our findings. Additionally, while our work focuses on unlearning methods designed for privacy-centric applications, such as removing personal data, it may not fully address other unlearning needs, such as removing harmful content. Moreover, our benchmark mainly evaluates fine-tuning-based unlearning methods using sophisticated loss functions, leaving unexplored other broader unlearning techniques, such as analytical or mechanical approaches. Another challenge lies in the scalability of these unlearning methods, as they may struggle to scale efficiently when applied to larger models and datasets, hindering their potential use in real-world systems. Furthermore, our focus on catastrophic forgetting overlooks unintended side effects, such as the introduction of biases or the degradation of model performance on unrelated tasks, and the broader impact of unlearning on fairness and safety remains an open area for future research.

Ethics

We utilized 84 hours of A100 GPU computation for our experiments, which resulted in an estimated 9 kg of CO2 emissions.

References

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52.

Denis Bobkov, Vadim Titov, Aibek Alanov, and Dmitry Vetrov. 2024. The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9337–9346.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. *Preprint*, arXiv:2112.03570.

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael Abu-Ghazaleh, M. Salman Asif, Yue Dong, Amit K. Roy-Chowdhury, and Chengyu Song. 2024. Cross-modal safety alignment: Is textual unlearning all you need? *Preprint*, arXiv:2406.02575.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics.

Jiali Cheng and Hadi Amiri. 2023. Multidelete for multimodal machine unlearning.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

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

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. 2024. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *Preprint*, arXiv:2403.01218.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. 2024. Model sparsity can simplify machine unlearning. *Preprint*, arXiv:2304.04934.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. *Preprint*, arXiv:1912.04958.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. *Preprint*, arXiv:2302.09880.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, and Sheng Bi. 2024a. Single image unlearning: Efficient machine unlearning in multimodal large language models. *arXiv preprint arXiv:2405.12523*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. 2024b. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Preprint*, arXiv:2403.03218.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024c. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8640–8650.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. *Preprint*, arXiv:2203.12817.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

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

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252.

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. *Preprint*, arXiv:2407.06460.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Ayush Tarun, Vikram Chundawat, Murari Mandal, and Mohan Kankanhalli. 2021. Fast yet effective machine unlearning.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*.

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2020. Stylespace analysis: Disentangled controls for stylegan image generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12858–12867.

x.ai. 2024. Grok-1.5 vision preview.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *ArXiv*, abs/2402.09801.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. Large language model unlearning. *Preprint*, arXiv:2310.10683.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operations. *arXiv preprint arXiv:2306.14870*.

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

Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. 2020. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7731–7740.

A Unlearning Methods

This section describes the main unlearning approaches considered in this work.

1. **Finetuning on retain data.** The most straightforward method to conduct unlearning is to finetune the model on the retain set, assuming that the model will unlearn the knowledge from the forget set and preserve its performance on the retain set. Despite its simplicity and reasonable effectiveness for relatively small models, it is not usable in models with huge sizes of pre-train sets, such as most LLMs.

2. **Gradient ascent on forget set.** In this method, unlearning is done by maximizing the loss on forget data with the intuition that it will lead to getting predictions that are dissimilar from the correct answers for forget set and consequently unlearning desired information. Thus, this method can be considered as a finetuning procedure with the following loss function:

$$L(D_F, \theta) = \frac{1}{|D_F|} \sum_{x \in D_F} NLL(x, \theta),$$

where $\mathrm{NLL}(x, \theta)$ is the negative log-likelihood of the model on the input x.

Instead of maximizing the NLL loss, maximizing the entropy of the model's predictions on the forget set is possible. The intuition behind this trick is that it will correspond to the increase of the model's uncertainty in its predictions on forget set, which will also correspond to successful unlearning.

3. **Gradient difference.** (Liu et al., 2022) The next method builds on the concept of combining two previous methods. It aims to increase the loss on the forget data and at least maintain the loss on the retain set. The loss function is defined as follows:

$$L_{GD} = -L(D_F, \theta) + L(D_R, \theta),$$

where D_F is the forget set that remains constant, D_R is the retain set that is randomly sampled during training, and L is a suitable loss function.

4. **KL minimization** . This approach aims to minimize the Kullback-Leibler (KL) divergence between the model's predictions on the retain set before and after unlearning while maximizing the conventional loss on the forget set. The L_{KL} loss function is defined as

$$\frac{1}{|D_F|} \sum_{x \in D_F} \frac{1}{|s|} \sum_{i=2}^{|s|} \mathrm{KL}\left(P(s_{<i}|\theta) \middle\| P(s_{<i}|\theta')\right).$$

The total objective function is formulated as follows:

$$L_{obj} = -L(D_F, \theta) + L_{\mathrm{KL}},$$

where θ' is the model's weights before unlearning, s is the input sequence, L is conventional loss, and $P(s|\theta)$ is the model's logits on the input sequence s with weights θ.

5. **IDK tuning**. Introduced in (Maini et al., 2024), this method aims to minimize the loss on the retain set, meanwhile, it uses pairs of inputs and "I don't know"(or some variations) labels instead of the original labels on the forget set. The loss function is defined as follows:

$$L_{idk} = L(D_R, \theta) + L(D_F^{idk}, \theta),$$

where L is some loss function, D_R is retain set, and D_F^{idk} is forget set with labels replaced with "I don't know" answers or some variations of them.

6. **Preference Optimization**. Inspired by Direct Preference Optimization (DPO) (Rafailov et al., 2023), the unlearning task can be framed as a preference optimization problem. In DPO, the model is trained to optimize user preferences directly, typically by maximizing the alignment between the model's outputs and the user's desired outcomes. Similarly, the goal of unlearning can be viewed as removing specific knowledge or patterns that the model has learned, effectively optimizing the model's outputs to align with new preferences that exclude the undesired information.

In this context, the unlearning task aims to adjust the model's parameters such that the output reflects a change in the learned distribution, making the model "forget" specific

pieces of knowledge. This can be formalized as a preference optimization problem, where the preference is towards outputs that no longer rely on unwanted data. Let L represent the loss function used for this task, which balances the model's performance on new data and its ability to unlearn specific information.

A common approach is to use a loss function that minimizes the difference between the model's current predictions and the desired "unlearned" predictions of the chosen reference model. The following loss function was considered to optimize for unlearning:

$$L = \lambda_1 L_{\text{task}}(D_F^{idk}, \theta) + \lambda_2 L_{\text{DPO}}(\pi_\theta, \pi_{ref}),$$

$$\begin{aligned} &L_{DPO}(\pi_\theta, \pi_{ref}) = \\ &= -\mathop{\mathbb{E}_{x,y\in D_F}}_{y'\in D_F^{idk}} \Big[\log\sigma(\beta\log\frac{\pi_\theta(y'|x)}{\pi_{ref}(y'|x)} - \\ &- \beta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)})\Big], \end{aligned}$$

where π_θ is related to the unlearned model which we try to optimize, σ is the sigmoid function, π_{ref} is reference model which in our case is fine-tuned on D_F^{idk} data, where labels are replaced with "I don't know" answers, (x, y) is input-answer pair from the forget set, y' is "I don't know"-like answer corresponding to this pair, $L_{\text{task}}(D_F^{idk}, \theta)$ is the standard task loss (e.g., cross-entropy) on the set D_F^{idk}, and $L_{\text{DPO}}(\pi_\theta, \pi_{ref})$ is DPO loss used for unlearning, which penalizes the model for retaining unwanted knowledge, computed between the input data x and the undesired in terms of unlearning labels y. λ_1 and λ_2 are weighting coefficients that balance the trade-off between task performance and the unlearning process (equal to 1 both), and β is the DPO coefficient (taken as 0.1 in our setting).

This formulation allows the model to optimize for maintaining task performance while ensuring the forgetting of specified information, similar to the dual objective in preference optimization. In the same way that DPO tailors the model to user preferences, this method shapes the model to "prefer" forgetting certain information, effectively unlearning it.

7. **Negative Preference Optimization** . Proposed in (Zhang et al., 2024) this method can be treated as DPO without positive examples. In our setting, the final loss function L_{NPO} for this method is derived as follows:

$$\frac{2}{\beta}\mathbb{E}_{x,y\in D_F}\left[\log\left(1+\left(\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right)^\beta\right)\right],$$

where all the notation is the same as for the previous DPO method. β was also taken equal to 1. Such loss functions ensure that the model output probability $\pi_\theta(y|x)$ is as small as possible, corresponding to the unlearning objective of the forget data.

8. **Teacher-Student (SCRUB)** (Kurmanji et al., 2023) The main idea of this method is to train a student model, which is taken as a desired unlearned model from the original one, such that it will "disobey" the teacher original model on the forget set. The resulting loss of student model in this method is constructed as follows:

$$d(x, w^s) = \text{KL}(p(f(x; w^o))||p(f(x; w^s))),$$

$$L_R = \frac{\alpha}{|D_R|}\sum_{x_r\in D_R} d(x_r, w^s),$$

$$L_F = \frac{1}{|D_F|}\sum_{x_f\in D_F} d(x_f, w^s),$$

$$L_{\text{task}} = \frac{\gamma}{|D_R|}\sum_{x_r\in D_R} l(x_r, y_r),$$

$$L = L_R - L_F + L_{\text{task}},$$

where $f(x; w^o)$ is the original teacher model with weights w^o, which are kept unchanged, $f(x; w^s)$ is the unlearned student model with parameters w^s, which are optimized, $d(x, w^s)$ is the KL-divergence between the output distributions of the student and teacher models on the input x, ℓ is the conventional task loss (e. g. cross-entropy), and α and γ are the hyperparameters controlling the importance of the student model's performance on the retain set. In our setting, α and γ were both set to 1. By minimizing this final loss L, the student model is expected to improve its performance on the retained set while unlearning from the forgotten set, respectively.

9. **LLMU** (Yao et al., 2024b)

   This method was proposed in one of the first works on unlearning LLMs (Yao et al., 2024b). In our experiments, we made slight modifications to the original method, and employed the following loss function:

$$
\begin{aligned}
&L_F := -L(D_F, \theta), \\
&L_r := \sum_{(x_F, y_r) \in D_F \times Y_r} \frac{1}{|y_r|} L(x_F, y_r, \theta), \\
&L_R := \sum_{x, y \in D_R} \mathrm{KL}(p_\theta(y|x) || p_{\theta'}(y|x)), \\
&L_{LLMU} = L_F + L_r + L_R,
\end{aligned}
$$

   where θ is the vector of unlearned model parameters, and θ' is the vector of original model parameters. This loss consists of three parts. The first one, L_F, is the negative conventional loss on the forget set, the optimization of which corresponds to the unlearning of the forget set. The second part, L_r, is the loss associated with "I don't know" labels (the original method used randomly generated labels), which also reinforces the forgetting of the D_F set. The third part is the KL divergence between the model's predictions on the retain set before and after unlearning, and its optimization relates to preserving the model performance on the retain set D_R. Note that it uses forward KL divergence instead of the usual reverse KL divergence.

10. **Representation Misdirection for Unlearning (RMU).** (Li et al., 2024b) This method builds on the thesis that the model's intermediate activations contain its knowledge about current inputs. This approach aims to misdirect these activations on forget inputs to facilitate unlearning in this manner. The loss for this method has the following form:

$$
\begin{aligned}
&L_{\mathrm{F}} = \mathbb{E}_{x \in D_F} \left[\frac{1}{|x|} \sum_{t \in x} ||h(t) - c \cdot u||_2^2 \right], \\
&L_{\mathrm{R}} = \mathbb{E}_{x \in D_R} \left[\frac{1}{|x|} \sum_{t \in x} ||h(t) - h_o(t)||_2^2 \right], \\
&L_{\mathrm{RMU}} = L_{\mathrm{F}} + L_{\mathrm{R}},
\end{aligned}
$$

   where $h(t)$ are the unlearned model's (which weights are optimized during unlearning procedure) hidden states on specific layer ℓ on input t, $h_o(t)$ are the hidden states of the original model (which parameters are frozen) on the layer ℓ on input t, u is the unit random vector with independent elements sampled uniformly from $[0, 1)$, and u kept fixed throughout unlearning, and c and α are hyperparameters controlling activations scaling and tradeoff between forgetting the D_F and retaining D_R respectively. The intuition behind this loss is to make the model's outputs on forget set D_F as far as possible from the correct ones by making hidden states as close as possible to random ones due to L_F summand and then build the outputs upon this states while making the final model closer to original one on the retain set with the help of L_R part of the loss. ℓ was chosen equal to 7 according to the empirical recommendation from the original method paper.

11. **Twins.** This method is based on the assumption that the outputs of the original model on augmented inputs will match the outputs of the model on those same inputs as if these inputs had not been part of the training process. The advantage of this method lies in the fact that it does not rely on a min-max optimization problem, which ensures its stability. However, a drawback is that this method is not applicable if the model was trained with augmentations. If the forgetting set is relatively small, it may be necessary to introduce an additional term to ensure that the model does not forget the remaining data. In this case, the loss function can be formulated as follows:

$$
\begin{aligned}
&L_{\mathrm{F}} = d(f(x_F), f_o(x_F^{aug})) \\
&L_{\mathrm{R}} = d(f(x_R), f_o(x_R)), \\
&L = L_{\mathrm{F}} + L_R,
\end{aligned}
$$

   where $d(a, b)$ represents the distance between vectors a and b, which can be either the L2 norm or KL divergence, $f(x)$ denotes the output of the unlearned model for input x. In contrast, $f_o(x)$ refers to the output of the original frozen model on the input x.

12. **SCRUB$_{bio}$.** This method adapts the original SCRUB for biometric task. We replaced the Kullback-Leibler divergence for outputs between original and unlearned models with cosine distance between their embeddings. Con-

sequently, the loss function for the task is formulated as follows:

$$L_{\mathrm{F}} = \frac{1}{|D_F|} \sum_{x_f \in D_F} \left(1 - d_{cos}(f(x_f), f_o(x_f))\right),$$

$$L_{\mathrm{R}} = \frac{1}{|D_R|} \sum_{x_r \in D_R} d_{cos}(f(x_r), f_o(x_r)),$$

$$L = L_{\mathrm{F}} + L_R,$$

where $d_{cos}(a, b)$ is the cosine distance between vectors a and b, $f(x)$ is the output of the unlearned model on input x, $f_o(x)$ is the output of the original frozen model on the input x.

13. **Sparsity** (Jia et al., 2024) This method is based on finetuning the model on the retain set using L1-regularization. The final loss is as follows:

$$L = L_{\mathrm{R}} + \lambda \cdot ||\theta||_1,$$

where λ is a parameter of regularization.

14. **Gradient Orhogonalization** This method maximizes the loss of the original task on the forget set D_F by ascending in the tangent direction of the gradient of the loss on the retain set ∇L_R. The resulting weight update step is as follows:

$$\theta_{i+1} = \theta_i + \eta \left(\nabla L_F - \frac{(\nabla L_F, \nabla L_R)}{|\nabla L_R|} \nabla L_R \right)$$

where $(\cdot, \cdot)$ is the scalar product, and η is the learning rate. This method requires a very small learning rate and many unlearning epochs due to the instability and the complexity of convergence. In our experiments, we used 400 unlearning epochs followed by 100 epochs of finetuning. As shown in Figure 5, the effects of 400 unlearning epochs were effectively undone by just one epoch of subsequent finetuning on retain set D_R.

B The process of face generation

To generate a set of the author's faces, we used StyleGAN 2 ADA (Karras et al., 2020). Using the generator, we synthesized a batch of 32 faces from

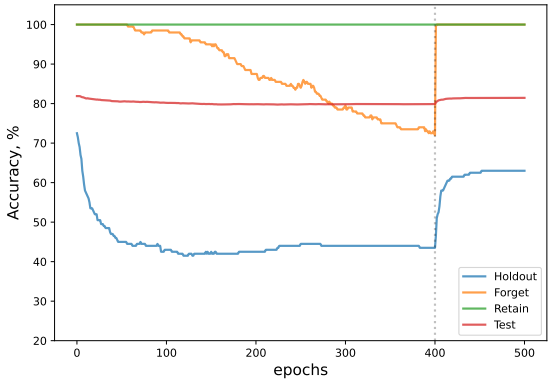

Figure 5: Process of unlearning with tangent gradient maximization. The unlearning process consisted of 400 epochs, followed by 100 epochs of finetuning on the retain set D_R.

the randomly sampled $z \in \mathcal{N}(0, I)$. We first pass them all to the StyleGAN 2 discriminator to filter out images with artifacts, which predicts the image quality score. We select only eight images with the best scores and discard the others. This process is repeated until 2000 images are collected.

We first synthesize a bath of 32 random faces to generate a set of older people. For each of them, we apply StyleFeatureEditor (Bobkov et al., 2024) with editing direction "age" from (Shen et al., 2020) and editing power 5, which increases the person's age. However, we noticed that this edit often adds glasses that shift the faces' distribution. To eliminate this effect, we also use StyleFeatureEditor after increasing age: we apply editing direction "glasses" from (Wu et al., 2020) with edit power -10. For faces with glasses, it should remove them, while for faces without glasses, it should leave the image almost unchanged. Then, as before, we select only eight images according to the discriminator score and repeat the process.

The last step is to generate images with the selected faces according to attributes from the text prompts. For this purpose, we used the personalized generation diffusion model PhotoMaker V2 (Li et al., 2024c). **According to our request, GPT-4o has generated prompts in such a way that the first sentence of a prompt describes the person, and the other sentences describe the setting, style, atmosphere, pose, and so on.** PhotoMaker requires a particular input type with the trigger word "img" and a particular class word (e.g., man, child or person) before it. For this purpose, we replaced the first sentences as follows: "a real photo of a {*old*} {*gender*} called {*name*} img, showing face." where *old* is "old" if the person is older than

60, "otherwise; *gender* is "man" or "woman" according to the person's gender, and *name* is the person's name. Below is an example of such a prompt:

"a real photo of an old man called Jaime Vasquez img, showing his face. Include his birth date, February 25, 1958, subtly in the background. The setting should reflect elements of the time period, such as vintage clothing styles or a retro ambience. Jaime should be depicted in a neutral pose, focusing on his character and era, with a hint of true crime elements around him."

To increase the power of the prompt, we used style strength = 0.5 and guidance scale = 7.5. We also used the same negative prompt "(asymmetry, worst quality, low quality, illustration, 3d, 2d, painting, cartoons, sketch), open mouth" for all images. The number of sampling steps was set to 50. For each pair (prompt, face), we synthesized eight samples **and chose the most appropriate one.**

C A sample of dataset

Our dataset consists of 200 fictitious authors, each with 15-20 visual and 20 textual questions. We add an example of data for a single person in the Table 5.

D Textual-only unlearning hyperparameters

For unlearning of the textual domain only, we use the textual part of CLEARbenchmark, containing question-answer pairs of about 200 authors, 20 for each of them (4000 pairs in total), and use the splits of size 90% and 10% of the entire data for retain and forget parts respectively. The "Gold" model for the further unlearning quality evaluation is trained on the retain data only, conducting 5 epochs of training with the batch size of 4, 1 gradient accumulation step, learning rate of 10^{-5}, weight decay of 0.01, and also applying LoRA adapter with the rank 8, $\alpha = 32$ and 0 dropout parameter. For the unlearning, we first finetune the model on the entire data split with the same hyperparameters: 5 epochs of training, batch size of 4, 1 gradient accumulation step, learning rate of 10^{-5}, weight decay of 0.01, LoRA rank of 8, $\alpha = 32$, 0 dropout coefficient. Then, unlearning methods are conducted on the forget data with the following hyperparameters: 5 epochs of unlearning, batch size of 4, 1 gradient accumulation step, learning rate of 10^{-5}, weight decay of 0.01, LoRA rank of 8, $\alpha = 32$, zero probability dropout. Such experimental settings and hyperparameters are the same for both Llama2-7B and Mistral architectures. To assess the unlearning quality, we compare the obtained unlearned model with the "gold" one and calculate **ROUGE-L** on retain and forget parts, **Forget Quality** and **Model Utility** metrics.

E CV pipeline

In this study, we evaluate each unlearning method from two key perspectives: its similarity to the gold standard (retraining from scratch) and its forgetting efficacy (error on the forget set). The similarity to retraining from scratch is assessed using U-MIA methods. Following the methodology of (Hayes et al., 2024), we employ population U-MIA and per-example U-LIRA.

We begin by taking a ResNet-18 pretrained on ImageNet and finetuning it for a biometric task using the Celeb dataset. We then train 256 ResNet-18 models using stochastic gradient descent (SGD) on a randomly selected half of the visual portion of our dataset, comprising 100 identities. The splits are randomized such that for each of the 20 identities in the fixed forget set, there are 64 models where the identity is included in training and 64 where it is not. Training is conducted for 20 epochs using the SGD optimizer with a learning rate of 0.1, batch size of 256, and weight decay of 5e-5.

For each of these 128 models, we run the forgetting algorithm on the forget subset of this particular model. From the resulting 128 models, we randomly select 64 target models (the remaining 64 will be used as shadow models for U-MIA and U-LIRA methods, see section G) on which the quality of the forgetting algorithms will be tested. Each of the 64 target models forgets a sample $\mathcal{D}f$ of 20 personalities. Additionally, for each target model, we form a holdout set D_H by selecting 20 personalities that were not used in the training of this model.

In our experiments, we employ U-LIRA with 64 shadow models, with half representing the in-distribution and the other half representing the out-distribution for each target example. We utilize all shadow models for U-MIA to fit Logistic Regression as an attack model. Both types of attacks use logits as input, which we compute for our biometric models as follows:

$$l = \log\left(\frac{\max(0, \cos(v, v_{enroll}))}{1 - \max(0, \cos(v, v_{enroll}))}\right),$$

where v represents the embedding of the target example x, ensuring $v = f(x)$, v_{enroll} denotes the enrolled vector for the corresponding individual, calculated as the mean of the embeddings from several supporting images of that particular identity, given by $v_{enroll} = \frac{1}{n}\sum_{i}^{n} f(x_i)$. In our studies, we use $n = 5$. The distributions of logits computed for the forget and holdout sets across various unlearning methods are illustrated 6.

F Multimodal unlearning hyperparameters

In a multimodal setting, we use both visual and textual parts of CLEARdataset, which consists of 4000 textual pairs of questions and answers about 200 authors, 20 for each of them, and 3770 images related to corresponding authors (number of images is less than the number of pairs because of GPT guard breaks and bugs in TOFU benchmark, as was described above). Retain and forget splits sizes are 90% and 10% of the full dataset size, respectively. The "Gold" model is trained on the retain data only with 3 epochs of training, batch size of 12, 1 gradient accumulation step, learning rate of 10^{-5}, weight decay of 0.01, LoRA rank of 8, $\alpha = 32$ and 0 dropout parameter. Unlearned models are also first finetuned on the full dataset with the same hyperparameters: 3 epochs of training, batch size of 12, 1 gradient accumulation step, learning rate of 10^{-5}, weight decay of 0.01, LoRA rank of 8, $\alpha = 32$, 0 dropout parameter. After that, unlearning techniques are applied to the model on the forget data using the following hyperparameters: 5 epochs of unlearning, batch size of 1, 2 gradient accumulation steps, learning rate of 10^{-5}, weight decay of 0.01, LoRA rank of 8, $\alpha = 32$, 0 dropout coefficient. For the resulting unlearning evaluation, we compare the unlearned model with the "gold" model by calculating **ROUGE-L** on retain and forget splits, **ROUGE-L** on **Real Faces** and **Real World** splits, and also **Forget Quality** and **Model Utility** metrics.

G U-MIA and U-LIRA

In this section, we provide details on evaluating unlearning methods using Unlearning Membership Inference Attack (U-MIA) algorithms. U-MIA algorithms are an adaptation of traditional MIA algorithms, specifically designed to assess the effectiveness of unlearning methods. The primary distinction between standard MIA and its unlearning counterpart lies in their objectives. Traditional MIA algorithms aim to determine whether a particular example was included in the training dataset of a model. In contrast, U-MIA algorithms are designed to detect whether a model was initially trained on a specific example and then subjected to an unlearning algorithm or if the model has never encountered the example at all.

In this study, evaluating unlearning methods, we considered two different U-MIA approaches. The first one is based on the original MIA introduced in (Shokri et al., 2017). It assumes training a specific classifier which for any input example (x, y) will output the probability that object x was forgotten by the model. The second one exploits the LIRA approach introduced in (Carlini et al., 2022). It is based on the Likelihood-ratio Test between hypotheses H1 and H2, where H1: object x comes from Q1 (forget distribution) and H2: x comes from Q2 (holdout distribution).

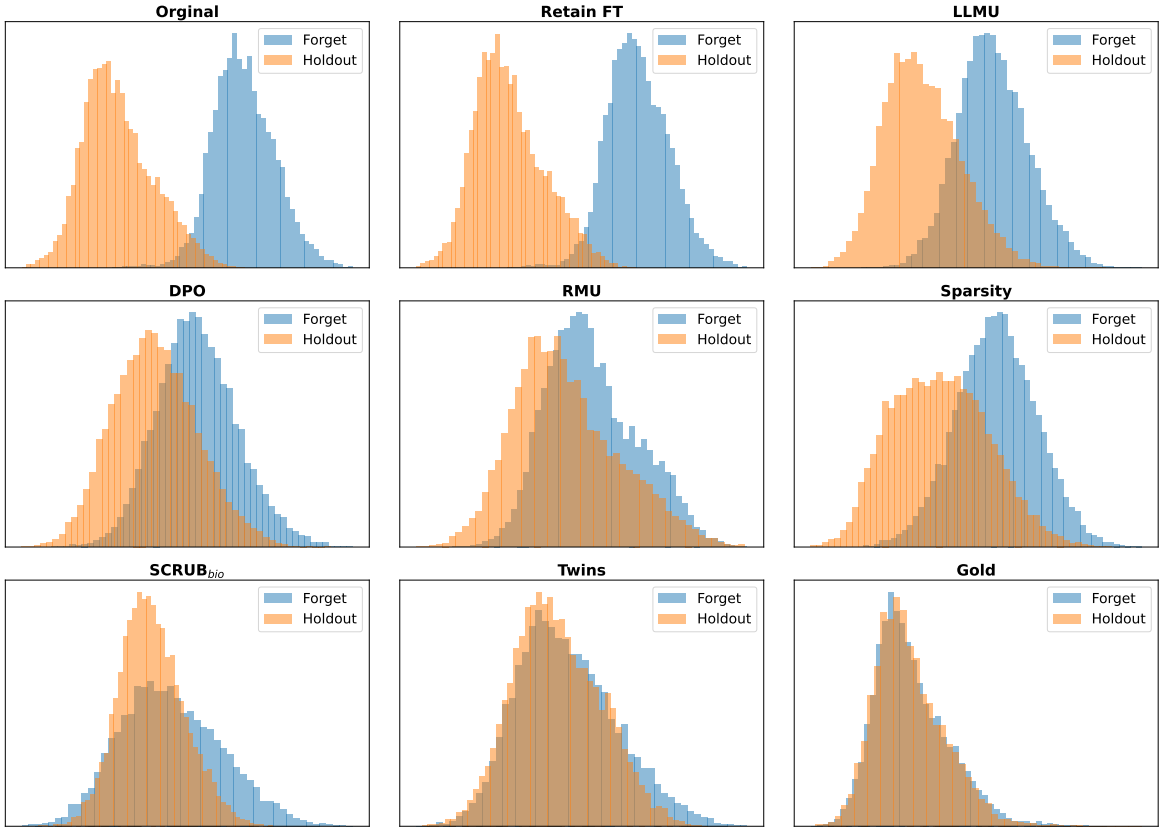

Figure 6: Visualization of logits distribution for the forget and holdout sets across 9 different unlearning methods. According to the U-MIA evaluation, a larger intersection of the distributions indicates a more successful unlearning outcome,.

Image	Caption
	Chukwu Akabueze in a striped shirt with a fleur-de-lis pin, looking directly at the camera in a vintage setting with a calendar in the background.
	Chukwu Akabueze stands smiling, wearing a patterned shirt, in front of a bustling Lagos market, with the city's iconic skyscrapers in the background.
	Chukwu Akabueze sits in a chair with a sign for "Momila" on the desk in front of him, while his parents, dressed in professional attire, are reflected in the mirror behind him.
	Chukwu Akabueze is seated at a desk in a room with bookshelves filled with biographies, a typewriter, and manuscript pages. He's smiling and looking directly at the camera.
	Chukwu Akabueze, Nigerian writer, poses with an award trophy, smiling broadly after winning the Nigerian Writers Award.
	Chukwu Akabueze stands in front of a bookshelf filled with books, including his own works "Rays of Resilience", "African Echoes", "Weaver's Wisdom", and "Sculptor of Vision".
	Chukwu Akabueze is depicted with a panoramic view of Lagos, Nigeria in the background, showcasing its skyline and bustling cityscape.
	Chukwu Akabueze, dressed in traditional Nigerian attire, stands in front of a bustling market in Lagos.
	Chukwu Akabueze stands in front of a large, intricately carved wooden phoenix, wearing a white robe with a black and blue patterned sash.
	Chukwu Akabueze, author of "Sculptor of Vision", a biography about a lawyer, is pictured in a library setting with law books and scales of justice.

Table 5: An example of all image-name pairs related to a single person