

# 000 001 DR.PO: DUAL REFERENCE AND PREFERENCE OP- 002 TIMIZATION FOR MACHINE UNLEARNING IN LARGE 003 LANGUAGE MODEL 004 005

006 **Anonymous authors**  
007 Paper under double-blind review  
008  
009  
010

## 011 ABSTRACT 012

013 Most high-performing LLM unlearning methods use preference learning but have  
014 a critical flaw: insufficient or singular positive preferences make post-unlearning  
015 models generate meaningless, inconsistent, or single-category outputs. These differ in probability  
016 distribution from "fully unlearned" models, causing suboptimal  
017 unlearning quality and privacy risks of unlearned data. We propose DR.PO (Dual  
018 Reference and Preference Optimization). It adopts answers with incorrect facts  
019 and answers indicating information deficiency as positive preferences, with  
020 distinct reference models for each. Instead of random sampling the full dataset for  
021 the retain set (as in existing methods), we match each forget sample to a corre-  
022 sponding retain sample via similarity scores—reducing repeated sampling of the  
023 retain set. Experiments show our method not only achieves better unlearning qual-  
024 ity and better privacy protection but also effectively preserves the model's original  
025 capabilities.

## 026 1 INTRODUCTION 027

028 Since deep learning models require the use of massive amounts of data during the training pro-  
029 cess, concerns have been raised by relevant personnel regarding data privacy, copyright, and secu-  
030 rity (Mantelero, 2013; de la Torre, 2018; Wang et al., 2024). Consequently, machine unlearning  
031 for LLMs has attracted significant attention from researchers. In previous studies, researchers have  
032 explored techniques such as gradient ascent, token-level perturbation, and preference learning to im-  
033 plement machine unlearning in LLMs, achieving certain results. Among these, preference learning  
034 methods often demonstrate superior performance.

035 However, due to the absence or singularity of positive preferences in existing preference learning  
036 methods, there exists a discrepancy between the output distribution of the unlearned model when  
037 exposed to the forgotten data and that of a model that never learned the forgotten data in real scenar-  
038 ios. This makes it challenging to strike an effective balance between unlearning quality and privacy  
039 protection of the forgotten data. To address this, we propose **Dual Reference and Preference Opti-  
040 mization (DR.PO)**, a preference learning method that employs dual positive preferences paired  
041 with corresponding reference models. By leveraging dual positive preferences to guide the model's  
042 unlearning process, DR.PO enables the model to produce an output distribution closer to that of  
043 exact unlearning when encountering the forget data.

044 Additionally, when examining existing methods, we observed that to maintain model performance,  
045 these methods typically incorporate randomly sampled samples from the retention set in each train-  
046 ing epoch. We argue that this process is problematic in terms of computational resource consumption  
047 and stability, and may amplify the impact of certain retention set samples that are less affected by the  
048 unlearning process. Therefore, we instead use a fixed set of top-k retention samples with the highest  
049 similarity to the forgotten data in the training process, and control the influence of these retention  
050 samples based on their similarity scores.

051 In summary, our main contributions are as follows:  
052

- 053 • **Dual Positive Preference:** We propose a dual positive preference forgetting method. Through the  
synergy of dual positive preferences, answers indicating missing information, and answers with

054 incorrect facts, the model can accommodate both preferences simultaneously in its outputs, thus  
 055 better protecting the privacy of forgotten data.  
 056

- 057 • **Dual Reference Model:** We propose using the pre-finetuning model as the reference model for the  
 058 positive preference of missing information. Experiments show that this better-aligned selection of  
 059 the reference model can better exert the effect of preference guidance.
- 060 • **Similarity-scores Retain:** We propose to select the retained data participating in unlearning based  
 061 on vector similarity, and use vector similarity to restrict their degree of influence. Experiments  
 062 show that this method can achieve performance close to that of the original random sampling and  
 063 is more stable.

## 064 2 RELATED WORK AND PRELIMINARIES

### 065 2.1 HOW CAN WE STATE LLM UNLEARNING?

066 Given an LLM  $\pi_\theta$  with parameters  $\theta$  and a dataset  $D = \{(x_i, y_i)\}_{i=1}^n$ , where  $(x_i, y_i)$  means an  
 067 question-answer pair in  $D$ , we denote the forget set as  $D_f \subset D$ , the retain set as  $D_r = D - D_f$ , the  
 068 LLM that has been trained on  $D$  as  $\pi_{full}$  and the pretrained LLM has not been trained on  $D$  as  $\pi_{base}$ .  
 069 The goal of LLM Unlearning is to make the  $\pi_{full}$  "forget" the specific information contained in  $D_f$   
 070 - i.e., to transform  $\pi_{full}$  into  $\pi_{unl}$  such that the performance of  $\pi_{unl}$  is approximately consistent  
 071 with those of  $\pi_{ret}$  only trained on  $D_r$ . Specially,  $\pi_{unl}$  should not retain dependencies on  $D_f$  and  
 072 ensure the unlearning process only eliminates the influence of  $D_f$  without distorting the model's  
 073 utility derived from  $D_r$  and other datasets.  
 074

### 075 2.2 WHAT ARE THE EXISTING LLM UNLEARNING METHODS?

076 Due to the massive parameter size of deep learning models, the method of achieving exact unlearning  
 077 through retraining with a retain set faces the challenge of high computational costs in practical  
 078 scenarios. Consequently, researchers tend to focus on developing a series of approximate unlearning  
 079 methods, which fine-tune model parameters using a forget set. Such approaches have already  
 080 yielded beneficial results in fields including federated learning (Halimi et al., 2022; Dhasade et al.,  
 081 2023; Gu et al., 2024; Jin et al., 2023), lifelong learning (Du et al., 2019; Parisi et al., 2019), im-  
 082 age generation (Li et al., 2024a; Zhang et al., 2024c; Gandikota et al., 2023; Zhang et al., 2024b),  
 083 graph neural networks (Li et al., 2024c; Cheng et al., 2023; Tan et al., 2024), and recommendation  
 084 systems (Sinha et al., 2025; Li et al., 2023).  
 085

086 In complex LLM Unlearning tasks, researchers have proposed a series of benchmark tests. These  
 087 include WMDP, which focuses on the unlearning of hazardous knowledge (Li et al., 2024b); TOFU,  
 088 which centers on personal privacy information (Maini et al., 2024); and MUSE, which addresses the  
 089 privacy/copyright issues of data owners (Shi et al., 2025).  
 090

091 Methodologically, some researchers have attempted Gradient Ascent (Maini et al., 2024) and its  
 092 derived methods, namely Gradient Difference (Liu et al., 2022) and WGA (Wang et al., 2025),  
 093 based on the idea of Gradient Ascent; some have explored token-level unlearning methods, such as  
 094 UNDIAL (Dong et al., 2025), SatImp (Yang et al., 2025), and PDU (Entesari et al., 2025); others  
 095 have proposed RMU (Li et al., 2024b), which perturbs model parameters related to the data to  
 096 be unlearned. Although the aforementioned methods that directly perform unlearning on model  
 097 parameters have shown certain effects, their performance in terms of unlearning efficacy, model  
 098 performance, and privacy protection of the unlearned data is unsatisfactory, leading to inconsistent  
 099 or incoherent responses generated by the model.  
 100

101 Therefore, some other researchers have started from preference learning and designed a series of  
 102 methods based on DPO (Rafailov et al., 2023). ( $\beta$  in the following equations is the regularization  
 103 strength)

104 **NPO** (Zhang et al., 2024a) retains only the negative preference component and guides the model to  
 105 avoid generating correct responses:

$$106 L_{NPO} = -\frac{2}{\beta} \log \sigma(-\beta \log \frac{\pi_\theta(y_f|x_f)}{\pi_{full}(y_f|x_f)}) \quad (1)$$

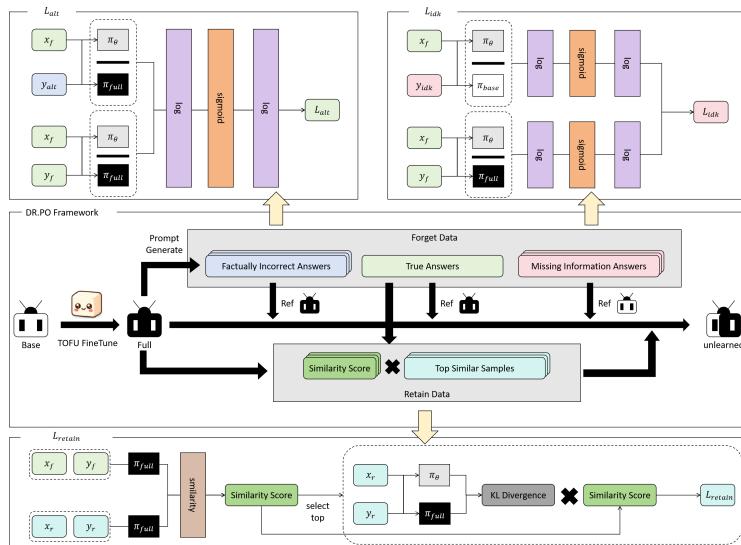


Figure 1: The framework of DR.PO.

**SimNPO** (Fan et al., 2024) is improved based on NPO: it removes the reliance on reference models and sets up a reward margin control to avoid uneven allocation of optimization resources( $\gamma$  is the reward margin parameter):

$$L_{SimNPO} = \mathbb{E} \left[ -\frac{2}{\beta} \log \sigma \left( -\frac{\beta}{|y_f|} \log \pi_\theta(y_f | x_f) - \gamma \right) \right] \quad (2)$$

**IdkPO** (Maini et al., 2024) uses answers indicating missing information as positive preference responses, guiding the model to be more inclined to indicate missing information when dealing with unlearned data( $y_{idk}$  is a response randomly selected from a number of pre-prepared answers indicating missing information):

$$L_{IdkPO} = -\frac{2}{\beta} \log \sigma \left( \beta \log \frac{\pi_\theta(y_{idk} | x_f)}{\pi_{full}(y_{idk} | x_f)} - \beta \log \frac{\pi_\theta(y_f | x_f)}{\pi_{full}(y_f | x_f)} \right) \quad (3)$$

**AltPO** (Mekala et al., 2025) uses prompts to require the model to generate a set of answers with incorrect facts based on the correct answers to forget data. It takes these answers with incorrect facts as positive preference responses and guides the model to provide factually incorrect answers( $y_{alt}$  is the generated factually incorrect answers):

$$L_{AltPO} = \mathbb{E}_{y_{alt}} \left[ -\frac{2}{\beta} \log \sigma \left( \beta \log \frac{\pi_\theta(y_{alt} | x_f)}{\pi_{full}(y_{alt} | x_f)} - \beta \log \frac{\pi_\theta(y_f | x_f)}{\pi_{full}(y_f | x_f)} \right) \right] \quad (4)$$

### 3 METHODS

Figure 1 shows the overview of our method.

#### 3.1 DR.PO: DUAL REFERENCE AND PREFERENCE OPTIMIZATION

Although Preference Optimization (PO)-type methods have achieved better unlearning performance, due to the way positive preferences are selected, they often perform poorly in terms of the privacy of forget data. Specifically, IdkPO only selects missing information answers as positive preferences, and AltPO only selects factually incorrect responses as positive preferences. All these issues result in a tendency toward a single response pattern.

For the reasons as mentioned above, and considering the fact that LLMs randomly generate either answers indicating missing information or answers with incorrect fact when addressing unknown

162 questions—with the former typically being what humans expect and require the model to output  
 163 during fine-tuning (He et al., 2025), and the latter arising from the hallucination phenomenon of  
 164 LLMs—we propose a method that can simultaneously integrate the positive preferences of IdkPO  
 165 and AltPO, namely **DR.PO: Dual Reference and Preference Optimization**.

166 Building on the integration of IdkPO and AltPO losses (Maini et al., 2024; Mekala et al., 2025),  
 167 DR.PO modifies the loss function of IdkPO. Since the effectiveness of Preference Optimization  
 168 (PO)-type methods depends to some extent on the choice of reference model, and PO is conceptually  
 169 based on responses that the model itself could generate with a certain probability, it is clearly  
 170 more appropriate to use a pre-trained model  $\pi_{base}$  that has not been exposed to the unlearned data as  
 171 the reference model for responses indicating missing information. Meanwhile, for correct responses,  
 172 the model  $\pi_{full}$  is still used as the reference model. Furthermore, due to the use of different reference  
 173 models for positive and negative preferences, the original loss form cannot be mathematically  
 174 retained; instead, we adopt two single-preference losses as substitutes.

$$L_{idk} = \mathbb{E}_{y_{idk}} \left[ -\frac{2}{\beta} \left( \log \sigma \left( \beta \log \frac{\pi_\theta(y_{idk}|x_f)}{\pi_{base}(y_{idk}|x_f)} \right) + \log \sigma \left( -\beta \log \frac{\pi_\theta(y_f|x_f)}{\pi_{full}(y_f|x_f)} \right) \right) \right] \quad (5)$$

$$L_{alt} = \mathbb{E}_{y_{alt}} \left[ -\frac{2}{\beta} \log \sigma \left( \beta \log \frac{\pi_\theta(y_{alt}|x_f)}{\pi_{full}(y_{alt}|x_f)} \right) - \beta \log \frac{\pi_\theta(y_f|x_f)}{\pi_{full}(y_f|x_f)} \right] \quad (6)$$

$$L_{DR.PO} = L_{idk} + L_{alt} \quad (7)$$

182 where  $\beta$  is the regularization strength, as in other PO methods.

### 183 3.2 RETAIN DATA PROCESSING BASED ON KL AND VECTOR SIMILARITY

185 Most LLM Unlearning methods involve fine-tuning with part of the retained dataset. Typically,  
 186 random sampling is performed on the complete retained dataset, and a subset of the retained dataset  
 187 is used for training. This approach can be traced back to Gradient Difference (Liu et al., 2022), which  
 188 adds the log-likelihood loss on a random subset of the retained dataset to the loss of the unlearned  
 189 dataset—a practice that has been adopted consistently by subsequent methods (Zhang et al., 2024a;  
 190 Maini et al., 2024; Mekala et al., 2025).

191 Although this approach to using the retained dataset has been proven effective on benchmarks, we  
 192 still have concerns: random sampling means that the coverage is completely random. While it may  
 193 achieve effective coverage for small datasets, for large datasets, it will likely require sampling more  
 194 samples to maintain model performance, thereby increasing computational costs. Based on our  
 195 intuition, we propose selecting only retained dataset samples with high similarity to the unlearned  
 196 samples for training. We use vector similarity as a coefficient and utilize these retained dataset  
 197 samples in the form of KL divergence (Maini et al., 2024) to constrain the similarity of responses  
 198 on the retained dataset between the model undergoing unlearning and the non-unlearned model,  
 199 which we have proven effective.

$$L_{retain} = \sum_{i=1}^k sim(\pi_{full}(x_r^{(i)}, y_r^{(i)}), \pi_{full}(x_f, y_f)) KL(\pi_{full}(y_r^{(i)}|x_r^{(i)}) || \pi_\theta(y_r^{(i)}|x_r^{(i)})) \quad (8)$$

$$L_{DR.PO-with-retain} = L_{idk} + L_{alt} + L_{retain} \quad (9)$$

205 where  $k$  represents the number of top- $k$  retained data set samples selected based on the highest sim-  
 206 ilarity to the unlearned samples.  $sim$  computes the vector similarity between two distributions; we  
 207 use the vector of the last token from the output of the model’s final hidden layer as the feature vector  
 208 for the input question-answer pair.  $KL$  measures the divergence between the output distributions of  
 209 the two models.

## 210 4 EXPERIMENTS

### 211 4.1 BENCHMARK

215 Proposed in 2024, TOFU is currently the most commonly used LLM Unlearning benchmark, on  
 which most existing methods have been evaluated. It mainly includes information on 200 fictional

Data Type	TOFU 1%	TOFU 5%	TOFU 10%
Forget data	40(2 authors)	200(10 authors)	400(20 authors)
Retain data	3960(198 authors)	3800(190 authors)	3600(180 authors)
All Finetune data		4000(200 authors)	

Table 1: The Information of TOFU.

authors, sets three difficulty levels (1%, 5%, and 10% unlearning)(see in Table 1), and provides a variety of model evaluations. We selected TOFU as the benchmark for our experiments and chose Llama3.2-1B as the experimental model based on computational resources. (Maini et al., 2024)

Meanwhile, we used OpenUnlearning as our evaluation framework and the implementation framework for baselines. Proposed in 2025, OpenUnlearning is an LLM Unlearning framework that includes complete evaluation functionality for the TOFU Benchmark and provides implementations of a series of baseline methods. (Dorna et al., 2025)

## 4.2 METRICS

Based on the four aspects that need to be considered in LLM Unlearning - accuracy of unlearning, knowledge preservation, privacy protection training, and fluency in dialogue generation - we selected four metrics, namely Forget Quality, Model Utility, Privacy Leakage, and Forget Gibberish - for evaluation (Maini et al., 2024; Shi et al., 2025; Dorna et al., 2025).

**Forget Quality(FQ):** The metric proposed by TOFU for measuring forget quality compares the truth ratio distributions of two models through the Kolmogorov-Smirnov Test (KS-Test), and quantifies forget quality using the p-value of the test: a higher p-value indicates a smaller difference between the distributions of the two models, thus a better forgetting effect. (Maini et al., 2024)

**Model Utility(MU):** The metric proposed by TOFU to measure the extent to which a model retains performance on non-unlearned data aggregates three indicators—probability, ROUGE-L, and truth ratio—across three datasets (Retain Set, Real Authors, and World Facts) using the harmonic mean. (Maini et al., 2024)

**Privacy Leakage(Priv.):** MUSE proposes that although most existing unlearning algorithms can reduce verbatim or knowledge memorization, they generally suffer from severe privacy leakage—either “over-unlearning” leading to abnormal behaviors or “under-unlearning” retaining training traces. Therefore, this metric is designed to measure the degree of privacy leakage: a value greater than 0 indicates “over-unlearning,” while a value less than 0 indicates “under-unlearning.” (Shi et al., 2025; Hayes et al., 2024)

**Forget Gibberish(Gibb.):** OpenUnlearning proposed this metric to measure whether a model exhibits the phenomenon of “gibberish” on the forget set after unlearning. OpenUnlearning uses the “autonlp-Gibberish-Detector-492513457” model to evaluate outputs, where a lower score indicates a higher likelihood that the output text is “gibberish”. (Dorna et al., 2025; Jindal, 2021)

## 4.3 IMPLEMENTATION DETAILS

All our experiments were conducted on a single NVIDIA A800 GPU. For  $y_{alt}$  in  $L_{alt}$ , we used the generation script of AltPO to generate factually incorrect answers in the default recommended quantity of AltPO, i.e., five answers. Meanwhile, for  $y_{idk}$  in  $L_{idk}$ , we selected five answers from the 100 optional answers provided by IdkPO in each epoch. For the baselines, we configured the experimental settings by referring to the recommended parameters preset in the corresponding papers and open-unlearning. Details can be found in the appendix.

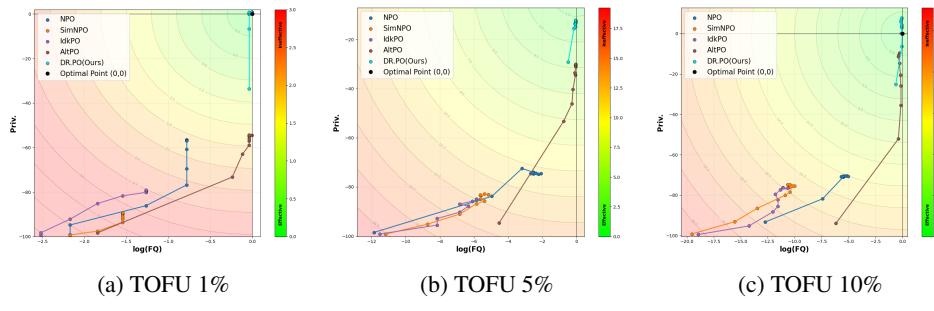
## 4.4 RESULTS

### 4.4.1 COMPARATIVE EXPERIMENT

Table 2 presents the experimental results of our method and other existing methods on the TOFU Benchmark under three settings: 1%, 5%, and 10%. From the results, it can be observed that methods of the PO type are more likely to achieve superior performance compared to other methods. In particular, AlterPO has basically reached the SOTA level, while other methods (Grad Ascent,

270 Table 2: Performance of various unlearning methods of TOFU 1%, 5% and 10% on Llama3.2-1B.  
 271 We use ( $\uparrow$ ) to indicate that a higher metric value is better, and ( $\rightarrow 0$ ) to indicate that a metric value  
 272 closer to 0 is better. In the "FQ" and "Priv." columns, the optimal results are highlighted in bold, and  
 273 the suboptimal results are underlined; in the "MU" column, results that are close to those of "Full"  
 274 and "Retain" are highlighted in bold, indicating no impairment caused by unlearning.

Method	TOFU 1%			TOFU 5%			TOFU 10%					
	FQ ( $\uparrow$ )	MU ( $\uparrow$ )	Priv. ( $\rightarrow 0$ )	Gibb. ( $\uparrow$ )	FQ ( $\uparrow$ )	MU ( $\uparrow$ )	Priv. ( $\rightarrow 0$ )	Gibb. ( $\uparrow$ )	FQ ( $\uparrow$ )	MU ( $\uparrow$ )	Priv. ( $\rightarrow 0$ )	Gibb. ( $\uparrow$ )
Full	6.76e-03	0.5992	-100.0000	0.8944	1.43e-12	0.5992	-99.9922	0.8584	3.91e-22	0.5991	-99.4574	0.8606
Retain	1.00e+00	0.5986	0.0000	0.8739	1.00e+00	0.5991	0.0000	0.9045	1.00e+00	0.5911	0.0000	0.9043
Grad Ascent	2.86e-02	<b>0.5903</b>	-59.5041	<b>0.8979</b>	1.94e-119	0.0000	-31.6578	0.1030	1.06e-239	0.0000	-17.8960	0.3702
Grad Difference (Liu et al., 2022)	2.86e-02	0.5206	-62.1015	0.8183	8.05e-07	0.4497	-13.1744	0.7453	1.60e-12	0.4365	20.2932	0.4159
WGA (Wang et al., 2025)	2.86e-02	<b>0.5993</b>	-86.8123	<b>0.8847</b>	2.96e-05	<b>0.5952</b>	-42.8227	<b>0.8475</b>	4.46e-06	<b>0.5984</b>	<b>3.2658</b>	0.7086
UNDIAL (Dong et al., 2025)	1.43e-02	<b>0.6097</b>	-83.8253	0.6561	6.57e-12	<b>0.6090</b>	-96.6692	0.7469	2.21e-19	<b>0.5973</b>	-97.0055	0.6709
SatImp (Yang et al., 2025)	1.43e-02	<b>0.6023</b>	-98.7485	<b>0.9184</b>	1.21e-10	<b>0.5965</b>	-97.6357	<b>0.8977</b>	1.18e-17	<b>0.5996</b>	-97.1979	<b>0.8889</b>
RMU (Li et al., 2024b)	5.41e-02	0.5595	-49.4687	0.7212	8.78e-02	0.5743	21.8240	0.5488	7.75e-04	<b>0.5831</b>	59.0735	0.0386
PDU (Entesari et al., 2025)	6.76e-03	<b>0.6036</b>	-100.0000	<b>0.8744</b>	5.951e-11	<b>0.5974</b>	-98.5296	0.8095	1.371e-07	<b>0.5837</b>	-40.1460	0.5656
NPO (Zhang et al., 2024a)	1.65e-01	<b>0.5977</b>	-48.4061	<b>0.9202</b>	4.31e-03	<b>0.5939</b>	-68.7696	<b>0.9016</b>	3.09e-06	<b>0.5957</b>	-70.5473	<b>0.8895</b>
SimNPO (Fan et al., 2024)	2.86e-02	<b>0.5982</b>	-83.2349	<b>0.9028</b>	2.38e-06	<b>0.5918</b>	-78.8347	<b>0.8834</b>	3.65e-11	<b>0.5954</b>	-74.8132	<b>0.8711</b>
IdkPO (Maini et al., 2024)	5.41e-02	<b>0.5837</b>	-72.7273	<b>0.9016</b>	1.39e-06	<b>0.5901</b>	-81.1598	<b>0.9082</b>	7.83e-12	<b>0.5905</b>	-76.2882	<b>0.9030</b>
AltPO (Mekala et al., 2025)	9.19e-01	<b>0.5952</b>	-54.3329	<b>0.8580</b>	<b>9.24e-01</b>	<b>0.5891</b>	-30.0012	<b>0.8460</b>	5.23e-01	<b>0.5844</b>	-9.6435	<b>0.8528</b>
DR.PO(Ours)	<b>9.90e-01</b>	<b>0.5954</b>	<b>0.9445</b>	<b>0.9315</b>	<u>8.66e-01</u>	<u>0.5960</u>	<u>-12.0491</u>	<u>0.8995</u>	<u>8.64e-01</u>	<b>0.6020</b>	<u>7.7646</u>	<b>0.8933</b>



298 Figure 2: The corresponding relationship between Forget Quality and Privacy Leakage during the  
 299 unlearning process of TOFU 1%, TOFU 5%, and TOFU 10% on Llama3.2-1B of various PO-type  
 300 unlearning methods.

303 Grad Difference, WGA, RMU, UNDIAL) tend to exhibit performance degradation relative to the  
 304 original model.

306 Across all three scales, our method achieves a better balance between Forget Quality and Privacy  
 307 Leakage. Specifically, under the 1% and 10% settings, our method reaches SOTA in both Forget  
 308 Quality and Privacy Leakage, demonstrating better forgetting effectiveness and privacy compared to  
 309 existing methods. Under the 5% setting, compared with AlterPO (which achieves SOTA in Forget  
 310 Quality), our method obtains significantly superior Privacy Leakage at the cost of slightly lower  
 311 Forget Quality—indicating that our method offers better privacy with comparable forgetting quality.

312 Meanwhile, compared to other methods, our method also maintains excellent performance in the  
 313 two metrics of Model Utility and Forget Gibberish, which shows that our method can effectively  
 314 avoid performance damage to the model caused by unlearning.

315 Figure 2 shows the combined variation of Forget Quality and Privacy Leakage during the unlearning  
 316 process of PO-type methods. It can be seen that as Forget Quality increases, Privacy Leakage will  
 317 also increase and eventually tend to be stable. The curves show that compared with other methods,  
 318 our method often has better Privacy Leakage at the same Forget Quality, which indicates that our  
 319 method has a better effect on the privacy protection of forgotten data than other methods.

320 In order to examine the impact of double positive feedback on the final answers generated by the  
 321 model, we additionally used the unlearned models obtained using various PO-type unlearning meth-  
 322 ods to generate responses for questions related to the forget data. We classified the generated answers  
 323 into categories using the Deepseek API: Irrelevant Answer, Missing Information Answer, or Fac-  
 324 tually Incorrect Answer. For each question related to the forget data, we instructed the model to

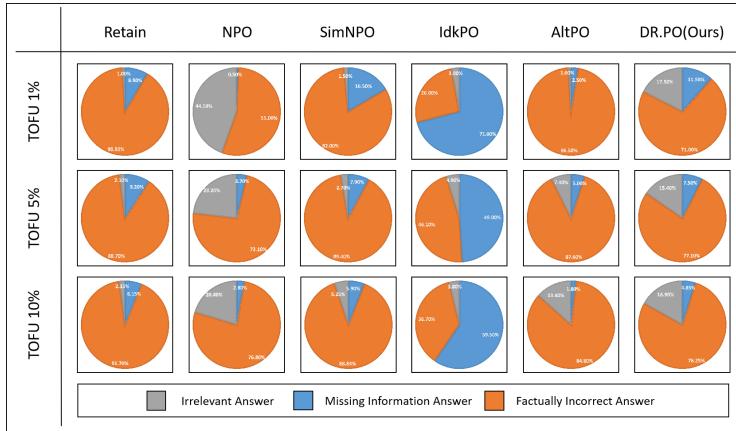


Figure 3: Regarding data forgetting questions, the proportion of various types of answers generated by different PO-type unlearning methods

generate 5 different responses and classify these answers using the Deepseek API (Liu et al., 2024; Guo et al., 2025). Figure 3 presents our experimental results. It can be observed that, compared to IdkPO and AltPO, our method is closer to the retain model in terms of the distribution of Missing Information Answers. However, our method produces more Irrelevant Answers, though it still outperforms NPO, which does not use positive preference. Meanwhile, the distribution of the three types of answers generated by our method is more stable compared to other methods—it does not exhibit significant changes due to fluctuations in the forgetting percentage, which is more consistent with the behavior of the retain model. This demonstrates that our method is indeed closer to the performance of the retain model when answering questions, and thus better protects the privacy of the forget data.

#### 4.4.2 ABLATION EXPERIMENT ON FORGET LOSS

To verify the effectiveness of the design of the forgetting loss in our method, we conducted ablation experiments on the forgetting loss component of our method. Our method mainly features the following aspects: (1) Adoption of a dual positive preference approach; (2) For answers indicating missing information, the model  $\pi_{base}$  without fine-tuning on  $D$  is used as the reference model; (3) For answers indicating missing information, a dual one-way preference superposition format is employed due to differences in reference models; (4) For answers indicating missing information, we adopt the same method of calculating expectations over multiple samples as that used for factually incorrect responses.

**Preference Ablation(no\_idk, no\_alt):** To verify Feature (1), we ablated  $L_{idk}$  and  $L_{alt}$  from the forgetting loss respectively and observed the effects.

**Reference Model Replacement(idk\_ref\_full):** To verify Feature (2), we replace  $L_{idk}$  with a form that, like  $L_{alt}$ , has dual preference outputs and both take  $\pi_{full}$  as the reference model. In this case,  $L_{idk}$  changes to:

$$L_{idk} = \mathbb{E}_{y_{idk}} \left[ -\frac{2}{\beta} \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_{idk}|x_f)}{\pi_{full}(y_{idk}|x_f)} - \beta \log \frac{\pi_{\theta}(y_f|x_f)}{\pi_{full}(y_f|x_f)} \right) \right] \quad (10)$$

**Don't Split Idk(idk\_no\_split):** To verify Feature (3), we retain the original form of the DPO loss and only modify the reference model in the positive preference part to  $\pi_{base}$ . At this point,  $L_{idk}$  is changed to:

$$L_{idk} = \mathbb{E}_{y_{idk}} \left[ -\frac{2}{\beta} \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_{idk}|x_f)}{\pi_{base}(y_{idk}|x_f)} - \beta \log \frac{\pi_{\theta}(y_f|x_f)}{\pi_{full}(y_f|x_f)} \right) \right] \quad (11)$$

**Use Only One Idk(idk\_single):** To verify Feature (4), we adopt the same approach as IdkPO: randomly selecting a single answer indicating missing information and applying it to  $L_{idk}$ , replacing the original form of sampling multiple answers and calculating the expectation.

378 Table 3: Performance of various forget loss ablation methods of TOFU 1% and TOFU 5% on  
 379 Llama3.2-1B.

381 Method	382 TOFU 1%			383 TOFU 5%			384 TOFU 10%					
	385 FQ (↑)	386 MU (↑)	387 Priv. (→ 0)	388 Gibb. (↑)	389 FQ (↑)	390 MU (↑)	391 Priv. (→ 0)	392 Gibb. (↑)	393 FQ (↑)	394 MU (↑)	395 Priv. (→ 0)	396 Gibb. (↑)
388 Full	6.76e-03	0.5992	-100.0000	0.8944	1.43e-12	0.5992	-99.9922	0.8584	3.91e-22	0.5991	-99.4574	0.8606
389 Retain	1.00e+00	0.5986	0.0000	0.8739	1.00e+00	0.5991	0.0000	0.9045	1.00e+00	0.5911	0.0000	0.9043
390 DR.PO	9.90e-01	0.5954	0.9445	0.9315	8.66e-01	0.5960	-12.0491	0.8995	8.64e-01	0.6020	7.7646	0.8933
391 DR.PO(no.idk)	9.19e-01	0.5979	-60.2125	0.8646	1.78e-01	0.5982	-59.4064	0.8957	3.22e-01	0.5978	-33.6984	0.8840
392 DR.PO(no.alt)	7.66e-01	0.5973	15.2302	0.9239	4.31e-03	0.5968	-17.5306	0.8848	1.78e-04	0.5941	20.2476	0.8799
393 DR.PO(idk_ref_full)	9.19e-01	0.5960	-44.9823	0.8748	1.63e-02	0.5992	-73.5492	0.9139	1.55e-01	0.6004	-49.1233	0.8718
394 DR.PO(idk_no_split)	9.19e-01	0.5953	-48.4061	0.8982	6.80e-02	0.5981	-69.5067	0.9014	7.83e-02	0.6009	-45.8525	0.8678
395 DR.PO(idk_single)	9.90e-01	0.5944	30.5785	0.9053	8.66e-01	0.5984	-9.1672	0.9007	7.00e-01	0.5955	3.0603	0.8814

389  
 390 Table 3 presents the experimental results of our ablation study on the forgetting set loss:

391  
 392 **Two positive preferences collaborate:** When either  $L_{idk}$  or  $L_{alt}$  is ablated, there is a significant  
 393 decline in both the forgetting quality and the privacy protection performance of the forgetting set.  
 394 This fully demonstrates the effectiveness and necessity of simultaneously using factually incorrect  
 395 responses and responses indicating information deficiency as positive preferences. Specifically, the  
 396 ablation of  $L_{idk}$  is more prominently manifested as a drop in privacy protection performance, while  
 397 the ablation of  $L_{alt}$  is more notably reflected in a decrease in forgetting performance. This indicates  
 398 that taking factually incorrect responses as positive preferences mainly dominates the quality of for-  
 399 getting, whereas using responses indicating information deficiency as positive preferences primarily  
 400 helps achieve the privacy protection of the forgetting set.

401  
 402 **It is necessary to match the reference model with the preference:** When  $L_{idk}$  is modified into  
 403 a form where both positive and negative preferences take  $\pi_{full}$  as the reference model, there is  
 404 also a significant decline in both the forgetting quality and the privacy protection performance of  
 405 the forgetting set. This indicates that using an inappropriate reference model for the corresponding  
 406 preferences can also affect the performance.

407  
 408 **When the reference models differ, the dual one-way preference outperforms the original DPO:** When  
 409  $L_{idk}$  is modified into the form of DPO loss and only the reference model in the positive pref-  
 410 erence part is adjusted, both the forgetting quality and the forgetting set privacy protection per-  
 411 formance of DR.PO also deteriorate. It can be seen that when the reference models differ, it is not a  
 412 good choice to adhere to the original DPO form rigidly; instead, using a dual one-way preference  
 413 for simulation can well replace the original DPO form.

414  
 415 **The positive preference should adopt the form of multiple expectations:** When a single sample  
 416 is randomly selected for the calculation of  $L_{idk}$ , the forgetting quality is mostly comparable to that  
 417 when multiple samples are randomly selected. Specifically, on TOFU 5% and TOFU 10%, there is  
 418 even a slight improvement in the privacy protection performance of the forgetting set. However, on  
 419 TOFU 1%, a distinct positive impulse signal emerges in the indicator for the forgetting set's privacy  
 420 protection performance. This indicates that when the data volume of the forgetting set is large,  
 421 the number of sampled items has no significant impact on performance, as a sufficient number of  
 422 positive preference samples can be learned. In contrast, when the data volume of the forgetting set  
 423 is small, over-unlearning is prone to occur due to the limited number of positive preference samples  
 424 available for learning.

#### 425 4.4.3 ABLATION EXPERIMENT ON RETAIN LOSS

426 Our method also incorporates minor designs for the retention loss. To verify this, we conducted  
 427 ablation experiments specifically on the retention loss, primarily to validate the effectiveness of two  
 428 modifications: using similarity scores as coefficients and adopting KL divergence to replace the  
 429 traditional NLL (Negative Log Likelihood) as the loss function, as well as the equivalence of using  
 430 the top-10 most similar retention set samples instead of randomly sampling retention set samples in  
 431 each iteration.

432 **no\_sim** indicates that similarity scores are not used as coefficients, meaning the coefficient for the  
 433 loss of all retention set samples is 1 when included in the calculation; **nll** denotes the use of tradi-  
 434 tional NLL (Negative Log Likelihood) to compute the loss of retention set samples; **random** refers

432 Table 4: Performance of various retain loss ablation methods of TOFU 1% and TOFU 5% on  
 433 Llama3.2-1B.

435 436 Method	TOFU 1%			TOFU 5%			TOFU 10%					
	FQ (↑)	MU (↑)	Priv. (→ 0)	Gibb. (↑)	FQ (↑)	MU (↑)	Priv. (→ 0)	Gibb. (↑)	FQ (↑)	MU (↑)	Priv. (→ 0)	Gibb. (↑)
437 Full	6.76e-03	0.5992	-100.0000	0.8944	1.43e-12	0.5992	-99.9922	0.8584	3.91e-22	0.5991	-99.4574	0.8606
438 Retain	1.00e+00	0.5986	0.0000	0.8739	1.00e+00	0.5991	0.0000	0.9045	1.00e+00	0.5911	0.0000	0.9043
439 DR.PO	9.90e-01	0.5954	0.9445	0.9315	8.66e-01	0.5960	-12.0491	0.8995	8.64e-01	0.6020	7.7646	0.8933
440 DR.PO(no_sim)	9.19e-01	0.5946	32.9398	0.8816	7.93e-01	0.5961	-13.2371	0.9000	3.67e-01	0.6005	16.6630	0.8737
441 DR.PO(null)	2.66e-01	0.5382	24.6753	0.8729	2.83e-04	0.4934	29.5483	0.8462	4.36e-09	0.4725	28.0162	0.8754
442 DR.PO(random)	9.90e-01	0.5952	-7.2019	0.9009	5.45e-01	0.5981	-25.6038	0.8772	7.00e-01	0.5999	2.0176	0.8786
443 DR.PO(retain01)	7.66e-01	0.5763	6.7296	0.9332	2.97e-02	0.5844	-56.6382	0.8933	2.29e-03	0.5863	-52.6837	0.8806
444 DR.PO(retain05)	9.19e-01	0.5918	0.4723	0.9285	3.94e-01	0.5908	-26.9683	0.8970	3.67e-01	0.5952	3.4035	0.8888
445 DR.PO(retain15)	9.19e-01	0.5958	4.2503	0.8770	7.95e-01	0.5945	-13.2411	0.8945	1.55e-01	0.5985	20.3003	0.8838

446 to the practice of sampling samples directly from the complete retention set in each epoch for retention  
 447 loss calculation, rather than using the pre-prepared top-10 samples with the highest similarity  
 448 scores to the forgetting set samples as the samples for retention loss calculation. Regarding the  
 449 number of samples selected from the retain set, in addition to the default of 10 samples, we also  
 attempted scenarios with 1, 5, and 15 samples, denoted as **retain01**, **retain05**, and **retain15**.

450 Table 4 presents our ablation results on  $L_{retain}$ . When replacing the KL divergence loss with the  
 451 conventional NLL loss, we observed a significant decline in Model Utility. This indicates that when  
 452 using a small amount of retain data highly correlated with the forget data for training, directly  
 453 applying KL divergence constraints yields better results in learning the corresponding probability  
 454 distribution compared to using NLL. When removing the similarity coefficient, Forgetting Quality  
 455 deteriorated. This phenomenon can be attributed to the fact that without the correlation coefficient,  
 456 all highly correlated retain data participated in training with equal intensity, thereby weakening the  
 457 forgetting effect. When replacing direct sampling of the top-most similar samples with random  
 458 sampling, our method maintained nearly consistent model performance. This demonstrates that  
 459 directly using highly correlated retain data can effectively replace the cumbersome and unstable  
 460 random sampling approach.

461 Table 4 also presents the results on different numbers of samples chosen from the retain set, re-  
 462 tain01 and retain05, which use fewer samples, exhibited significantly poor performance in terms of  
 463 unlearning quality. Additionally, retain01 performed poorly in the privacy protection performance  
 464 of the forget set data. This indicates that selecting fewer samples from the retain set can also affect  
 465 unlearning quality, and that the use of retain data during the unlearning process can influence the  
 466 model’s knowledge related to the forget data. retain15, which uses more samples, also showed a  
 467 decline in unlearning quality and generally exhibited over-unlearning in the privacy protection per-  
 468 formance of the forget set data. This suggests that selecting a larger number of samples from the  
 469 retain set can also affect the degree to which the unlearning process utilizes the forget set data.

## 470 5 CONCLUSION

471 Due to the large amount of data used by LLMs during training, potential issues related to privacy,  
 472 copyright, and data compliance have raised widespread concerns, sparking researchers’ interest in  
 473 LLM Unlearning. Among existing LLM unlearning methods, PO-type methods have demonstrated  
 474 superior performance in terms of unlearning quality. However, due to the lack of positive preference  
 475 or the single type of positive preference in existing methods, problems still exist in the forgotten  
 476 quality and privacy protection of forget data. Output distributions that differ from the retain model  
 477 may lead to privacy leakage of forget data. Based on this, we propose DR.PO, a method that employs  
 478 dual positive preference answers and pairs each with an appropriate reference model. A series of  
 479 experiments have demonstrated that our method exhibits superior performance in both unlearning  
 480 quality and privacy protection of forget data, and its output distribution is more consistent with that  
 481 of the retain model. Additionally, we noticed that the random sampling of retain data in each epoch  
 482 during training may cause waste of computational resources and instability. Therefore, we propose  
 483 a retain data sampling and loss calculation method based on vector similarity, and verify that it has  
 484 essentially equivalent performance to the original method while being more stable.

486 REFERENCES  
487

488 Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. Gnndelete: A general  
489 strategy for unlearning in graph neural networks. *arXiv preprint arXiv:2302.13406*, 2023.

490 Lydia de la Torre. A guide to the california consumer privacy act of 2018. *Available at SSRN*  
491 3275571, 2018.

492 Akash Dhasade, Yaohong Ding, Song Guo, Anne-marie Kermarrec, Martijn De Vos, and Leijie  
493 Wu. Quickdrop: Efficient federated unlearning by integrated dataset distillation. *arXiv preprint*  
494 *arXiv:2311.15603*, 2023.

495 Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL:  
496 Self-distillation with adjusted logits for robust unlearning in large language models. In Luis  
497 Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations*  
498 *of the Americas Chapter of the Association for Computational Linguistics: Human Language*  
499 *Technologies (Volume 1: Long Papers)*, pp. 8827–8840, Albuquerque, New Mexico, April 2025.  
500 Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.  
501 naacl-long.444. URL <https://aclanthology.org/2025.naacl-long.444/>.

502 Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico  
503 Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified bench-  
504 marking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025. URL <https://arxiv.org/abs/2506.12618>.

505 Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong anomaly detection  
506 through unlearning. In *Proceedings of the 2019 ACM SIGSAC conference on computer and com-*  
507 *munications security*, pp. 1283–1297, 2019.

508 Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Con-  
509 strained entropic unlearning: A primal-dual framework for large language models. *arXiv preprint*  
510 *arXiv:2506.05314*, 2025.

511 Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Sim-  
512 plicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Neurips*  
513 *Safe Generative AI Workshop 2024*, 2024. URL <https://openreview.net/forum?id=pVACX02m0p>.

514 Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts  
515 from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer*  
516 *vision*, pp. 2426–2436, 2023.

517 Hanlin Gu, Gongxi Zhu, Jie Zhang, Xinyuan Zhao, Yuxing Han, Lixin Fan, and Qiang Yang. Un-  
518 learning during learning: An efficient federated machine unlearning method. *arXiv preprint*  
519 *arXiv:2405.15474*, 2024.

520 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
521 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
522 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

523 Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning:  
524 How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

525 Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inex-  
526 act unlearning needs more careful evaluations to avoid a false sense of privacy, 2024. URL  
527 <https://arxiv.org/abs/2403.01218>.

528 Jianfeng He, Linlin Yu, Changbin Li, Runing Yang, Fanglan Chen, Kangshuo Li, Min Zhang, Shuo  
529 Lei, Xuchao Zhang, Mohammad Beigi, et al. Survey of uncertainty estimation in large language  
530 models-sources, methods, applications, and challenge. 2025.

531 Ruinan Jin, Minghui Chen, Qiong Zhang, and Xiaoxiao Li. Forgettable federated linear learning  
532 with certified data removal. *arXiv e-prints*, pp. arXiv–2306, 2023.

540 Madhur Jindal. Gibberish detector: High-accuracy text classification  
 541 model, 2021. URL [https://huggingface.co/madhurjindal/](https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457)  
 542 autonlp-Gibberish-Detector-492513457.

543 Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine unlearning for image-to-  
 544 image generative models. In *The Twelfth International Conference on Learning Representations*,  
 545 2024a. URL <https://openreview.net/forum?id=9hjVoPWPnh>.

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

557 Xunkai Li, Yulin Zhao, Zhengyu Wu, Wentao Zhang, Rong-Hua Li, and Guoren Wang. Towards  
 558 effective and general graph unlearning via mutual evolution. In *Proceedings of the AAAI Conference  
 559 on Artificial Intelligence*, volume 38, pp. 13682–13690, 2024c.

560 Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Yizhao Zhang, Zhongxuan Han, Dan Meng, and Jun  
 561 Wang. Making users indistinguishable: Attribute-wise unlearning in recommender systems. In  
 562 *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 984–994, 2023.

563 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 564 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint  
 565 arXiv:2412.19437*, 2024.

566 Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on  
 567 Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

568 Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task  
 569 of fictitious unlearning for llms, 2024. URL <https://arxiv.org/abs/2401.06121>.

570 Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the  
 571 ‘right to be forgotten’. *Computer Law & Security Review*, 29(3):229–235, 2013.

572 Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund  
 573 Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning fac-  
 574 tual knowledge in large language models. In Owen Rambow, Leo Wanner, Marianna Apidi-  
 575 anaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the  
 576 31st International Conference on Computational Linguistics*, pp. 3732–3752, Abu Dhabi, UAE,  
 577 January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.252/>.

578 German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual  
 579 lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

580 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea  
 581 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances  
 582 in neural information processing systems*, 36:53728–53741, 2023.

583 Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao  
 584 Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-  
 585 way evaluation for language models. In *The Thirteenth International Conference on Learning  
 586 Representations*, 2025. URL <https://openreview.net/forum?id=TArmA033BU>.

587 Yash Sinha, Murari Mandal, and Mohan Kankanhalli. Multi-modal recommendation unlearning for  
 588 legal, licensing, and modality constraints. In *Proceedings of the AAAI Conference on Artificial  
 589 Intelligence*, volume 39, pp. 12541–12549, 2025.

594 Jiajun Tan, Fei Sun, Ruichen Qiu, Du Su, and Huawei Shen. Unlink to unlearn: Simplifying edge  
 595 unlearning in gnns. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 489–492,  
 596 2024.

597 Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger.  
 598 Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth  
 599 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=huo8MqVH6t>.

600 Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive  
 601 survey. *arXiv preprint arXiv:2405.07406*, 2024.

602 Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring  
 603 criteria of loss reweighting to enhance llm unlearning, 2025. URL <https://arxiv.org/abs/2505.11953>.

604 Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catas-  
 605 trophic collapse to effective unlearning, 2024a. URL <https://arxiv.org/abs/2404.05868>.

606 Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia  
 607 Liu. Unlearnncanvas: A stylized image dataset to benchmark machine unlearning for diffusion  
 608 models. *CoRR*, 2024b.

609 Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong,  
 610 Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure  
 611 in diffusion models, 2024c. URL <https://arxiv.org/abs/2405.15234>.

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647