

# 000 001 002 003 004 005 CARE: CONTINUAL REAL-TIME UNLEARNING WITH 006 ENSURED PRESERVATION OF LLM KNOWLEDGE 007 008 009

010 **Anonymous authors**  
011 Paper under double-blind review  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053

## ABSTRACT

As concerns grow over the issue of large language models (LLMs) inadvertently internalizing sensitive or erroneous information, unlearning—the selective removal of undesired knowledge—has been drawing an increasing amount of attention. Existing approaches to unlearning fail to account for scenarios requiring immediate processing of knowledge removal requests, leaving services that rely on LLMs vulnerable to prolonged leakage of sensitive information while the process of unlearning is underway. Moreover, when such requests occur not just once, but continuously over the period of deployment, current methods cause LLMs to suffer increasingly degraded utility performance with the processing of each request. To address these issues, we propose **Continual Real-time Unlearning with Ensured Preservation of LLM Knowledge (CaRE)**. Prior to LLM deployment, we train an unlearning sentence embedder with a synthetically generated dataset designed to enable the formation of sharp decision boundaries for determining whether a given input query corresponds to any forget requests in the database. At inference, an embedding is generated for the input query and compared with the embedding of each forget request using a distance metric and the maximum score is compared to a threshold which is used to decide whether to answer the query or to refuse. Since our method does not modify any weights of the language model, it avoids catastrophic forgetting and is able to achieve near perfect knowledge preservation after an arbitrary number of updates. Our experiments on four benchmarks demonstrate that **CaRE** achieves a superior balance of forgetting and knowledge preservation over all existing methods in the continual setting while also being the only method capable of processing forget requests in real-time.

## 1 INTRODUCTION

With the rapid advancement of Large Language Models (LLMs), their applications have been swiftly expanding across society and into various aspects of daily life. However, many unforeseen challenges regarding their reliability are also coming to light (Ji et al., 2023; Chang et al., 2024; Gallegos et al., 2024; Zhao et al., 2024). One of these issues is the need for a way to reliably erase targeted pieces of information from an LLM in a localized manner. During the pre-training or finetuning of LLMs on large-scale datasets, they are at risk of incorporating and disseminating sensitive information (Carlini et al., 2021), including copyrighted or privacy-related content (Das et al., 2025), as well as incorrect knowledge (De Angelis et al., 2023). Not only does this raise significant ethical issues, but it also entails legal risk for developers of LLMs as ‘the Right to be Forgotten’ is mandated by regulations such as the General Data Protection Regulation (GDPR) (Mantelero, 2013) in the EU and the California Consumer Privacy Act (CCPA) (de la Torre, 2018) in the US.

As a countermeasure, LLM unlearning (Jang et al., 2022; Liu et al., 2025) has been introduced with the aim of efficiently removing inappropriate information while preserving the existing knowledge and capabilities of the model—without requiring full retraining. Most approaches to LLM unlearning (Zhang et al., 2024; Jia et al., 2024) apply a training algorithm to the LLM, utilizing a *forget set*—the data to be removed—and a *retain set*—the data used to preserve the model’s utility. The training objective is typically to either maximize the loss on the original input–output pairs of the *forget set* or minimize the loss on the same inputs paired with refusal responses, while also minimizing the loss on the *retain set* to preserve the model’s knowledge with respect to items that are not the subject of any forget requests.

054 However, existing approaches remain inadequate for deployment in real world scenarios for the fol-  
 055 lowing reasons. First, prior methods have mainly been designed and evaluated under the assumption  
 056 of a single, unchanging set of forget requests, and their performance is unlikely to hold up under  
 057 circumstances that require the sequential processing of a continual stream of new requests. In par-  
 058 ticular, most existing methods operate by modifying the weights of the target LLM, which impairs  
 059 its general ability and knowledge due to the phenomenon of catastrophic forgetting (Luo et al.,  
 060 2023), and this problem is only compounded by the accumulating forget requests in the continual  
 061 setting. Furthermore, while some prior studies have explored the continual setting, they still suffer  
 062 from catastrophic forgetting on non-target data, and exhibit poor generalization with respect to para-  
 063 phrased variants of questions in the *forget set*. This indicates that with existing continual unlearning  
 064 methods, LLMs not only experience utility degradation, but also fail to fully eliminate the informa-  
 065 tion specified in the forget requests. Finally, the nature of many such requests necessitates immediate  
 066 action to prevent further harm—as in the case of sensitive or dangerous information—while most  
 067 existing approaches rely on expensive and time-consuming optimization procedures applied to the  
 068 target LLM. In addition to the heavy cost in time and compute incurred by the training process it-  
 069 self, optimization-based methods typically entail some degree of hyper-parameter search (Bergstra  
 070 & Bengio, 2012) to find an acceptable balance between effectiveness of forgetting and preservation  
 071 of knowledge. In cases where a *retain set* is required, securing a sufficient quantity of high-quality  
 072 data fit for the task can cause even further delay (Gao et al., 2025).

073 To tackle these challenges, we introduce **Continual Real-time Unlearning with Ensured Preserva-  
 074 tion of LLM Knowledge (CaRE)**. Prior to LLM deployment, **CaRE** trains an unlearning sentence  
 075 embedder on a synthetically generated dataset with hard negatives designed to enable fine-grained  
 076 classification between user queries related to the *forget set*, and those that are unrelated. After the  
 077 LLM is deployed, **CaRE** continuously adds the embeddings of any received forget requests to its  
 078 embedding database in real-time and compares them with the embedding of the current user query.  
 079 Then based on this comparison, we decide whether the LLM should provide a response to the user  
 080 query or refuse. Importantly, since the unlearning embedder does not require any additional training  
 081 post-deployment—and in particular, does not need to use either the *forget set* or the *retain set* for  
 082 training, the entire process achieves significantly faster unlearning compared to prior approaches.  
 083 Moreover, because the weights of the LLM remain unmodified, **CaRE** allows for near perfect utility  
 084 preservation. As a result, not only does our method substantially outperform all other unlearning  
 085 methods in the continual setting (which is the setting most relevant to real world applications), it is  
 086 the first method we are aware of that is capable of processing ongoing forget requests in real-time  
 087 with minimal degradation of model performance as requests accumulate over time.

088 In summary, the contributions of our work are as follows:  
 089

- 090 • We introduce **CaRE**, an unlearning framework that entails virtually no overhead for processing  
 091 new forget requests and thus constitutes the first unlearning method capable of handling continual,  
 092 sequential forget requests in real-time.
- 093 • Through experiments across multiple benchmark datasets, we demonstrate that by leaving the  
 094 weights of the LLM unmodified, **CaRE** is able to largely circumvent the catastrophic forgetting  
 095 problem faced by existing methods and achieve near perfect preservation of LLM knowledge,  
 096 even after processing a long succession of continual forget requests.
- 097 • We demonstrate superiority over prior state-of-the-art (SOTA) unlearning methods in additional  
 098 aspects such as the ability of our method to generalize to any unlearning task after training on  
 099 a single dataset (whereas existing methods typically require retraining on every new *forget* and  
 100 *retain set*), and robustness to paraphrased variants of sentences in the *forget set*.

## 101 2 RELATED WORK

102 **Conventional Unlearning.** Methods that only use the *forget set* for training are called Gradient  
 103 Ascent (GA) (Jang et al., 2022). These methods train the target LLM to minimize a loss on the  
 104 *forget set* defined as the positive log likelihood of the text in the *forget set*, thereby minimizing the  
 105 likelihood of generating the information contained in the *forget set*. Other methods add to this loss by  
 106 including a term for the negative log likelihood of the text in the *retain set*, which acts as a regularizer  
 107 forcing the LLM to not only forget the information in the *forget set* but to also explicitly remember  
 108 the information in the *retain set*. These methods are known as Gradient Difference (GradDiff) (Liu

108 et al., 2022). A third approach, called Preference Optimization (PO) (Maini et al., 2024), uses a loss  
 109 that encompasses terms for both the *forget set* and the *retain set*, but instead of using the positive  
 110 log likelihood on the *forget set*, it uses the negative log likelihood on alternate refusal responses to  
 111 the questions in the *forget set*. Negative Preference Optimization (NPO) (Zhang et al., 2024) uses  
 112 the loss from Direct Preference Optimization (DPO) (Rafailov et al., 2023) but with only negative  
 113 examples (instead of pairs of positive and negative examples). More recent work includes SOUL (Jia  
 114 et al., 2024), which is not of itself a distinct unlearning method, but rather an improvement that adds  
 115 second-order optimization to existing methods. These methods tend to have weak performance on  
 116 knowledge preservation metrics as modifying weights inevitably results in catastrophic forgetting.  
 117

118 **Weight Preserving Unlearning.** Existing approaches that avoid modifying LLM weights include  
 119 In-Context Unlearning (ICUL) (Pawelczyk et al., 2023) which adds data points from the *forget*  
 120 set with perturbed labels as in-context examples to the LLM prompt, and guardrail methods (Thaker  
 121 et al., 2024) that add a filtering step by querying an auxiliary LLM to detect whether the output of the  
 122 target LLM is related to any data in the *forget set*. These methods generally have low performance  
 123 except for very large foundation models and they are not scalable as the increasing size of the *forget*  
 124 set will eventually cause issues due to context length limitations (Liu et al., 2023). Perhaps the  
 125 method that bears the greatest resemblance to our own is GUARD (Deng et al., 2025). This method  
 126 also trains a model to classify user queries as being either related or unrelated to the *forget set*.  
 127 However, the classifier they use is specific to the *forget set* it was trained on and thus needs to be  
 128 retrained for every new set of forget requests, which precludes the possibility of real-time unlearning  
 129 and makes it less suitable for the continual setting.

130 **Continual Unlearning.** Two methods that are particularly relevant to the present work are O3 (Gao  
 131 et al., 2025) and UniErase (Yu et al., 2025), both of which were designed specifically to address  
 132 unlearning in the continual setting. The former works by training an orthogonal low-rank adapter  
 133 (LoRA) (Hu et al., 2021) to unlearn the information in the *forget set*, and then trains an out-of-  
 134 distribution (OOD) detector to determine how much weight to give to the adapter during inference  
 135 based on how close the input query is to the data in the *forget set*. The latter method adds an  
 136 unlearning token “<UNL>” to the tokenizer vocabulary of the LLM and uses prompt tuning (Lester  
 137 et al., 2021) to train the model to output refusal responses whenever an input query is followed by  
 138 “<UNL>”. It then uses model editing methods (Meng et al., 2022) to modify the weights of the LLM  
 139 such that when questions from the *forget set* are input to the language model, it generates “<UNL>”  
 140 as the first token. As these methods both modify the weights of the target LLM (or its adapter), they  
 141 are still subject to the problem of catastrophic forgetting.

### 143 3 METHOD FOR REAL-TIME CONTINUAL UNLEARNING

144 To guarantee not only the preservation of LLMs’ existing capabilities, but also to enable effec-  
 145 tive real-time processing of successive forget requests in a continual unlearning setting, we pro-  
 146 pose **CaRE** (Continual Real-time Unlearning with Ensured Preservation of LLM Knowledge). Our  
 147 method begins by training an unlearning sentence embedder  $U$  that learns to generate embedding  
 148 vectors for user queries  $q$  and forget requests  $f$  whose distance can be used to form a decision bound-  
 149 ary for whether to answer the query or to refuse (Sec. 3.2). After the deployment of the LLM, we  
 150 perform real-time unlearning by asynchronously updating the *forget set* (via generating embeddings  
 151 for new forget requests  $U(f)$  as they are received and adding them to our embedding database) and  
 152 handling user queries to the LLM through our proposed inference pipeline in conjunction with the  
 153 trained  $U$  (Sec. 3.3). The overall framework of **CaRE** is illustrated in Figure 1.

#### 156 3.1 PROBLEM FORMULATION

157 To formalize our task, we begin by denoting  $D$  as the entirety of the data used to train the large  
 158 language model  $G$  that serves as the starting point for unlearning.  $D$  can be partitioned into two  
 159 splits, the forget split  $D_f$  and the retain split  $D_r$ , where the former represents all the data that needs  
 160 to be forgotten and  $D_r = D \setminus D_f$  represents the rest of the data, which needs to be preserved by the  
 161 language model. The gold standard of what we are trying to achieve with unlearning is a model  $G^*$

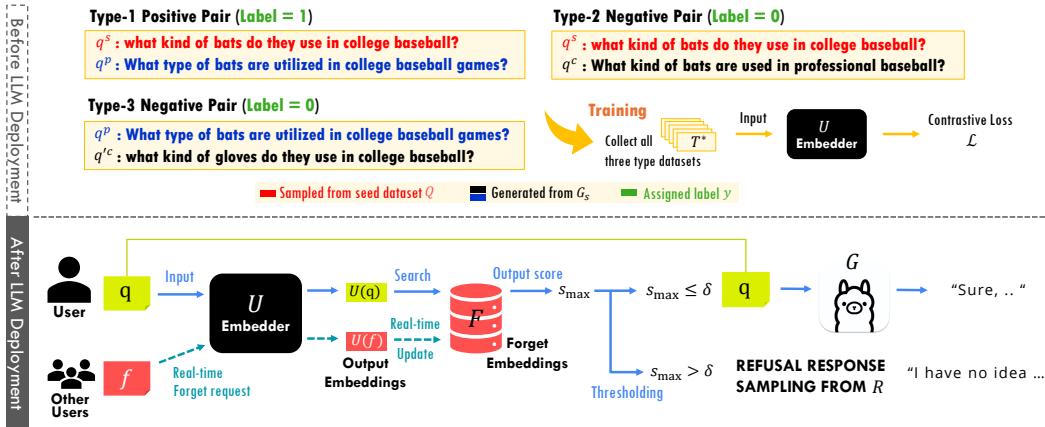


Figure 1: An overview of the **CaRE** framework. **CaRE** consists of a training phase carried out prior to deployment (upper part) and a three-step inference process after deployment (lower part). In the training phase, the embedder  $U$  is trained on three types of synthetic data generated from a seed dataset (training does not require any data from the *forget set* or *retain set*). For inference, real-time continual unlearning is enabled through three steps: (i) embed  $q$ , embed-and-store  $f$ , (ii) retrieval and thresholding, and (iii) decision on whether the LLM responds or refuses. Since the LLM’s weights remain unchanged, we are able to maintain a high level of utility preservation.

that has been trained in the same manner as  $G$ , but on  $D_r$  only. Such a model would not contain any knowledge of the data in  $D_f$  since it was never trained on  $D_f$  and it could be expected to contain roughly the same amount of knowledge about  $D_r$  as  $G$ , since it is assumed to have undergone the same training process on those data points.

In most real world applications,  $G^*$  is just a theoretical ideal that cannot be obtained in practice since modern LLMs are too large and costly to retrain from scratch. Hence, this objective is approximated by performance metrics on  $D_f$  that gauge how effectively the data in  $D_f$  has been forgotten and performance metrics on  $D_r$  that measure how well the rest of the data has been preserved. Most unlearning techniques involve modifying the weights of  $G$  to obtain an approximation  $\hat{G} \approx G^*$ , which subjects the language model to heavy drops in performance on  $D_r$  as the weight updates give rise to catastrophic forgetting (Luo et al., 2023), a problem that is worsened in the continual setting described below. Our method on the other hand, does not modify  $G$  at all, thus preserving its existing knowledge in tact and leaving the potential for achieving the same performance on  $D_r$  as  $G$  an open possibility.

**Continual setting.** To closer align our task with scenarios likely to be encountered in the real world, we additionally extend the unlearning task to the continual setting where the forget requests arrive successively and need to be processed cumulatively in sequence. Hence, we start with an initial partition  $D_{f_0}$ ,  $D_{r_0} = D \setminus D_{f_0}$  to which we apply our unlearning techniques and evaluate. Then the *forget set* is expanded to include new requests resulting in a new partition  $D_{f_1}$ ,  $D_{r_1} = D \setminus D_{f_1}$  such that  $D_{f_0} \subset D_{f_1}$  and we perform further unlearning on the same model to reflect the additional requests and evaluate once more. The goal is to maintain high performance on the forget and retain objectives over each stage until the final set of forget requests and final partition  $D_{f_N}$ ,  $D_{r_N} = D \setminus D_{f_N}$ . If finetuning is applied to  $G$  post-deployment to add new information,  $D$  itself may also expand, but for simplicity we assume that  $D$  is fixed.

Most existing unlearning methods use the entire forget split for training, hence the *forget set* used for training is simply  $D_f$ . Methods that also make use of the retain split for training cannot use the entire split since it is too vast, so they typically use a small subset consisting of counterexamples to the *forget set* which is termed the *retain set*  $D_{retain} \subset D_r$ . For evaluation, again typically the entire forget split  $D_f$  is used to test forgetting effectiveness, whereas to test preservation of knowledge, various subsets of  $D_r$  are used, including the *retain set* as well as utility datasets that are completely unrelated to the *forget set* to test general knowledge capacity, such as ‘‘World Facts’’ in the TOFU benchmark (Maini et al., 2024) and WinoGrande (Sakaguchi et al., 2019) in the RETURN benchmark (Liu et al., 2024).

216 3.2 PRE-DEPLOYMENT TRAINING  
217

218 We now describe the first step of the **CaRE** framework, which involves training the unlearning  
219 sentence embedder  $U$ . Before deployment of the large language model  $G$ , it is unknown what  
220 removal requests  $f$  may arise, or what queries  $q$  may be issued to  $G$ . Therefore,  $U$  must learn a  
221 representation that effectively distinguishes and generalizes over any possible future  $q$  and  $f$ , taking  
222 this uncertainty into account. It must also be robust to variations of the *forget set*, e.g., paraphrased  
223 sentences that convey the same information as those in the *forget set* should still trigger refusal to  
224 respond. To meet the above requirements, we build training data of three types through the following  
225 process.

226 First, we collect the questions from a seed QA dataset  $Q = \{q_1^s, q_2^s, \dots, q_n^s\}$ , e.g., Natural Questions  
227 (Kwiatkowski et al., 2019), where each  $q_i^s$  represents the question from the question-answer  
228 pair  $(q_i^s, a_i^s)$ . For each question  $q^s$ , we apply transformations as illustrated in Figure 1 to generate  
229 two variants of the question,  $G_s(\tau_1(q^s)) = (q^p, q^c)$ , where  $\tau_1(\cdot)$  is an input prompt template for a  
230 surrogate LLM  $G_s$ . Here  $q^p$  represents a paraphrased variant of  $q^s$  and  $q^c$  represents a contrastive  
231 variant.  $q^p$  is thus a rephrasing of  $q^s$  that should elicit the same response from the target LLM  $G$ .  
232 Coupled with  $q^s$ ,  $(q^s, q^p)$  constitutes a positive pair with label  $y^p = 1$ , which we term type-1 data.  
233 In contrast,  $q^c$  is a question designed to exhibit high lexical or syntactic overlap with  $q^s$  but differ in  
234 semantic meaning. Together with  $q^s$ , the pair  $(q^s, q^c)$  serves as a hard-negative example with label  
235  $y^c = 0$ , which we term type-2 data. Following the same procedure, we obtain the contrastive sample  
236 of  $q^p$  via  $\tau_2(\cdot)$ , denoted as  $q'^c = G_s(\tau_2(q^p))$ , which paired with  $q^p$  as  $(q^p, q'^c)$  forms an instance of  
237 type-3 data labeled with  $y'^c = 0$ , thereby functioning as an additional hard-negative sample along  
238 with the type-2 data. We apply the three types of data augmentation to every sample in  $Q$ , and  
239 construct the dataset  $T^* = \{[(q_i^s, q_i^p), y_i^p], [(q_i^s, q_i^c), y_i^c], [(q_i^p, q_i'^c), y_i'^c]\}_{i=1}^n$  for training the embedder  $U$ .  
240 We use  $T^*$  to finetune a pre-trained sentence embedding model (Reimers & Gurevych, 2019) using  
241 the following contrastive loss (Hadsell et al., 2006):

$$242 \mathcal{L}(T) = \frac{1}{2|T|} \sum_{(q, q', y) \in T} \left[ y \cdot d_U(q, q')^2 + (1 - y) \cdot \max(0, m - d_U(q, q'))^2 \right], \quad (1)$$

243 where  $d_U$  denotes a distance metric in the embedding space of  $U(\cdot)$ , which is the cosine distance in  
244 our case defined as  $d_U(q, q') = 1 - \frac{U(q) \cdot U(q')}{\|U(q)\| \|U(q')\|}$ .  $T \subset T^*$  is a batch of samples from the training  
245 dataset and  $m$  is an appropriately chosen margin. The loss serves to decrease the distance between  
246 positive examples and increase the distance between negative examples up to the margin  $m$ . The  
247 hard-negative samples in our dataset are designed to represent difficult edge cases, thereby enabling  
248 the embedder to form more fine-grained and precise decision boundaries in the embedding space.  
249 It should be noted that all of the above training is conducted without requiring either the *forget set*  
250 or *retain set*, and that it is carried out prior to the deployment of  $G$ . After deployment, the single  
251 trained  $U$  model can operate across any given forgetting task and domain without any additional  
252 training and its effectiveness is not limited to any particular *forget* and *retain set*.

253 3.3 POST-DEPLOYMENT INFERENCE  
254

255 Once  $G$  is deployed, **CaRE** performs unlearning and inference through the following three steps.  
256 **(i):** Given the  $m$ -th forget sample  $f_m$ , its embedding  $f_m^{\text{emb}} = U(f_m)$  is generated and stored in the  
257 set of forget embeddings  $F$ . The update of  $F$  is carried out immediately in real-time upon arrival of  
258  $f_m$  and can be expressed as

$$259 F = \{f_1^{\text{emb}}, f_2^{\text{emb}}, \dots, f_{m-1}^{\text{emb}}\} \Rightarrow F \leftarrow F \cup \{f_m^{\text{emb}}\}. \quad (2)$$

260 This instantaneous operation constitutes the entirety of our unlearning process post-deployment and  
261 stands in stark contrast to the heavy optimization procedures employed by other methods to unlearn  
262 a given set of forget requests. Asynchronously, whenever a user query  $q$  is input to  $G$ , it is projected  
263 into the embedding space as  $q^{\text{emb}} = U(q)$ . **(ii):** For each embedding  $f_i^{\text{emb}}$  in  $F$ , we compute  
264 its cosine similarity score  $s_i$  with respect to  $q^{\text{emb}}$ , and obtain the score set  $S = \{s_i\}_{i=1}^m$ , where  
265  $s_i \in [-1, 1]$ . Using  $S$ , we identify the element  $f_j \in F$  most related to  $q$  by taking an element with  
266 the maximum score  $s_j = s_{\max}$ , and check whether it exceeds a given threshold  $\delta$ . In this process,

270 the user queries sent to the LLM and requests for information removal are all handled continuously  
 271 and in real-time, without mutual interference. **(iii):** The final response  $r_{\text{res}}$  returned to the user is  
 272 defined as follows:

$$273 \quad r_{\text{res}} = \begin{cases} 274 \quad G(q), & \text{if } s_{\text{max}} < \delta, \\ 275 \quad \text{a sampled element from } R, & \text{if } s_{\text{max}} \geq \delta, \end{cases} \quad (3)$$

277 where  $R$  is a predefined set of refusal expressions such as “I don’t know” or “I can’t answer that  
 278 question”. If  $s_{\text{max}} < \delta$ , we determine that  $q$  is unrelated to any information in the current *forget set*,  
 279 and thus return the regular generated output for  $q$  using  $G$ . In contrast, if  $s_{\text{max}} \geq \delta$ , we determine  
 280 that  $q$  is closely related to some information in the *forget set* and therefore decline to answer  $q$ .  
 281 In this case, a refusal response is sampled from  $R$  and returned as  $r_{\text{res}}$  (Appendix F). Note that the  
 282 parameters of  $G$  are not modified at any step of this process. This guarantees knowledge preservation  
 283 within  $G$  thereby preventing the occurrence of catastrophic forgetting, which is key to our method  
 284 being able to maintain such high performance on the retain and utility datasets after processing an  
 285 arbitrary number of successive forget requests.

## 286 4 EXPERIMENTS

### 288 4.1 EXPERIMENTAL SETUP

290 **Benchmarks.** We conduct unlearning experiments in the continual setting using four widely used  
 291 benchmarks. **(1) Privacy Data Unlearning:** The RETURN benchmark (Liu et al., 2024) consists  
 292 of synthetically generated question-answer pairs related to real world individuals with Wikipedia  
 293 pages. The goal is to forget selected details (not all) about a subset of the individuals. We posit a  
 294 scenario where out of the 30 target individuals, three individuals issue forget requests at each stage,  
 295 resulting in a total of 10 stages of continual unlearning. **(2) General Science Knowledge Unlearn-**  
 296 **ing:** We adopt the setting in Gao et al. (2025) which uses a subset of the ScienceQA dataset (Lu  
 297 et al., 2022) as the *forget set* to sequentially unlearn four scientific topics: biology, physics, chemistry,  
 298 and economics. At each stage, one topic is added to the *forget set* and the remaining topics  
 299 make up the *retain set*. **(3) Fictitious Authors Unlearning:** TOFU (Maini et al., 2024) is an unlearn-  
 300 ing benchmark that fine-tunes a pre-trained language model on QA pairs about completely fabricated  
 301 authors to ensure that none of the data in the *forget set* exists in the pre-training data. The task is  
 302 then to unlearn information about a selection of the fake authors. We divide the authors into three  
 303 groups, resulting in a three-stage continual unlearning setup. **(4) False Information Unlearning:**  
 304 TruthfulQA (Lin et al., 2021) is a benchmark designed to assess whether LLMs provide factually  
 305 grounded answers to misleading questions across diverse topics (i.e., whether they avoid generating  
 306 misinformation). We adopt a continual unlearning setting in which all the questions are partitioned  
 307 into three stages and used as the *forget set*. Further details about the evaluation datasets can be found  
 308 in Appendix C.1

309 It should be noted that for the *forget set* used for evaluation, we replace the questions with para-  
 310 phrased variants as this is a more realistic assumption for real world use cases and using the same  
 311 questions verbatim from the original *forget set* would be trivial for our method to solve with 100%  
 312 accuracy by setting the decision boundary threshold  $\delta$  to 1. Also, for each benchmark we add a syn-  
 313 synthetically generated *near utility* dataset containing examples designed to be similar in appearance to  
 314 sentences in the *forget set*, but distinct in meaning (and hence should not be subject to removal—they  
 315 are edge cases designed to test the locality of the forgetting mechanism). The detailed procedure for  
 316 generating these datasets is outlined in Appendix E.

317 **Evaluation Metrics.** As our method does not modify any weights of the LLM, it does not alter  
 318 the probability distribution output by the LLM, which renders probability-based metrics such as the  
 319 Truth Ratio (Maini et al., 2024) meaningless for our case. Hence for most evaluation datasets we  
 320 use ROUGE-L (Lin, 2004) to measure the similarity between the generated response and the ground  
 321 truth answer. In cases where we are able to extract an exact answer from the generated response  
 322 using simple parsing, such as the WinoGrande dataset (Sakaguchi et al., 2019) and the ScienceQA  
 323 benchmark (Lu et al., 2022), we calculate accuracy using an exact match criterion.

324 **Baselines.** We selected GA (Jang et al., 2022), GradDiff (Liu et al., 2022), PO (Maini et al., 2024),  
 325 NPO (Zhang et al., 2024), SO-PO (Jia et al., 2024), GUARD (Deng et al., 2025), O3 (Gao et al.,

2025), and UniErase (Yu et al., 2025) as our baselines. Base indicates the target model prior to unlearning, which serves as an upper bound for knowledge preservation performance. UniErase only works on data given in (subject, relation, object) triplet form i.e. questions and answers about people, so we exclude it from our experiments for TruthfulQA and ScienceQA.

The training configuration of  $U$ , the details of  $\tau_1$  and  $\tau_2$ , and results for other models are presented in Appendices C.2, D, and H respectively.

## 4.2 PRIVACY DATA UNLEARNING

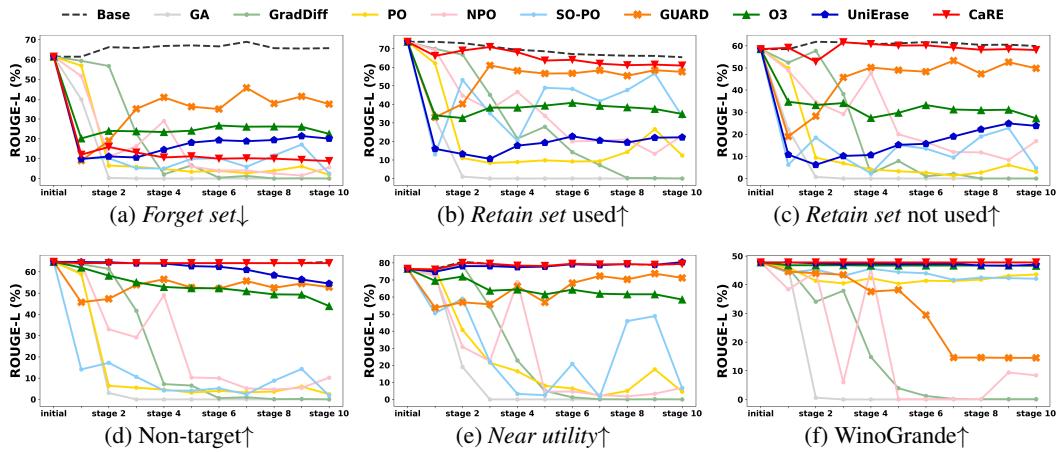


Figure 2: Continual unlearning results on RETURN. (a) indicates performance on the unlearning target, while (b)–(f) indicate performance on data that we aim to preserve (details in Appendix C.1).

Figure 2 presents our experimental results on the RETURN benchmark. The gradient-based and preference optimization methods exhibit a strong tendency towards overforgetting—they are successful in removing the knowledge related to the *forget set* but at the cost of significant degradation in performance on unrelated knowledge. We can clearly see a sharp drop-off from the base model as the stages progress—as expected due to catastrophic forgetting. GUARD, O3 and UniErase preserve knowledge to some extent, but fail to sufficiently remove the target knowledge. CaRE, on the other hand, achieves effective removal of the data from the *forget set* with negligible degradation in performance on the other datasets across all ten stages of evaluation.

## 4.3 GENERAL SCIENCE KNOWLEDGE UNLEARNING

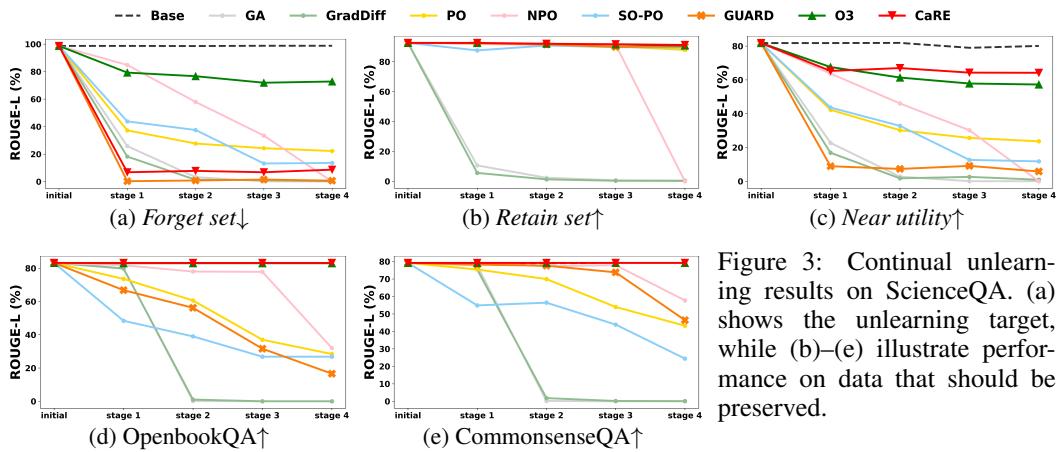


Figure 3: Continual unlearning results on ScienceQA. (a) shows the unlearning target, while (b)–(e) illustrate performance on data that should be preserved.

Figure 3 presents our results on the ScienceQA benchmark. The only method that is able to maintain comparable performance to CaRE on the knowledge preservation datasets across all stages of

378 evaluation is O3. However, we can see that its performance on the *forget set* is unusually poor. We  
 379 found that this is due to O3 being unable to generalize to paraphrased variants of the questions in  
 380 the *forget set*. While it is able to achieve much lower scores of 20.7%, 4.6%, 10.1%, and 11.8%  
 381 across the four stages of the original *forget set*, it is surprisingly brittle against even slight changes  
 382 in wording and thus cannot be said to have truly forgotten the information in the *forget set*. So  
 383 again **CaRE** is the only method able to achieve effective forgetting while maintaining near perfect  
 384 knowledge preservation across each stage of evaluation.

#### 385 4.4 FICTITIOUS AUTHORS UNLEARNING

388 Table 1: Results on the TOFU benchmark. **F.G.** (*forget set*), **R.T.** (*retain set*), **N.U.** (*near utility*),  
 389 **R.A.** (Real-Authors), and **W.F.** (World Facts) are reported; the best results are highlighted in **blue**,  
 390 and the second-best are underlined, excluding near-zero values on **F.G.** caused by over-forgetting.

Method	TOFU dataset for LLaMA2-7B-chat														
	Stage 1					Stage 2					Stage 3				
	F.G.↓	R.T.↑	N.U.↑	R.A.↑	W.F.↑	F.G.↓	R.T.↑	N.U.↑	R.A.↑	W.F.↑	F.G.↓	R.T.↑	N.U.↑	R.A.↑	W.F.↑
Base	0.496	0.973	0.620	0.940	0.913	0.518	0.973	0.617	0.940	0.913	0.509	0.973	0.599	0.940	0.913
GA	0.390	0.715	0.574	0.855	0.821	0.211	0.320	0.488	0.576	0.785	0.003	0.003	0.005	0.000	0.006
GradDiff	0.242	0.424	0.550	0.763	0.812	0.001	0.002	0.003	0.000	0.003	0.000	0.000	0.000	0.000	0.000
PO	0.110	0.873	0.598	0.923	0.883	0.111	0.801	0.533	0.692	0.862	0.181	0.860	0.570	0.897	0.877
NPO	0.072	0.874	<u>0.608</u>	<u>0.930</u>	0.892	<u>0.031</u>	0.796	0.601	0.912	0.900	0.065	0.815	0.593	0.914	0.895
SO-PO	0.094	0.837	0.586	0.899	0.896	0.118	0.808	0.592	0.922	0.868	0.120	0.791	0.562	0.916	0.873
GUARD	0.121	0.773	0.573	0.909	0.896	0.112	0.798	0.536	0.872	0.883	0.129	0.775	0.553	0.891	0.876
O3	0.128	0.338	0.564	0.651	0.905	0.070	0.093	0.198	0.095	0.282	0.083	0.093	0.163	0.079	0.219
UniErase	<u>0.047</u>	<u>0.947</u>	0.603	0.906	<u>0.930</u>	0.058	<u>0.943</u>	<u>0.610</u>	0.899	<u>0.930</u>	<u>0.062</u>	0.942	0.587	0.889	0.905
CaRE	<b>0.046</b>	<b>0.969</b>	<b>0.620</b>	<b>0.940</b>	0.913	0.055	<b>0.969</b>	<b>0.615</b>	<b>0.940</b>	0.913	<b>0.043</b>	<b>0.961</b>	<b>0.597</b>	<b>0.940</b>	<b>0.913</b>

401 Table 1 presents our results on the TOFU benchmark. The only method that appears to remain  
 402 competitive with our method across all three stages is UniErase. However, the apparent strength  
 403 of this method—which still lags **CaRE** in overall performance—should be weighed against the  
 404 inability of UniErase to handle any data that does not conform to its strict (subject, object, relation)  
 405 format, which is a significant limitation, as well as its inability to process forget requests in real-time.  
 406

#### 407 4.5 FALSE INFORMATION UNLEARNING

409 Table 2 reports the results for TruthfulQA. The objective in this case is  
 410 to prevent the dissemination of false  
 411 information contained in the *forget set*.  
 412 However, minimizing similarity  
 413 to a particular incorrect answer can  
 414 be gamed: the model may simply  
 415 produce a different incorrect response  
 416 while remaining untruthful. Hence,  
 417 instead of measuring the similarity of  
 418 the response to the answers in the *forget set*, we measure its similarity to a  
 419 set of refusal responses (the pairwise maximum from the set) such as “I don’t know” as our indica-  
 420 tion of success. This inherently restricts our evaluation to methods that are capable of optimizing  
 421 towards a desired response (i.e. it excludes gradient ascent methods that only optimize away from an  
 422 undesirable response). From the table we can see again that **CaRE** has much stronger performance  
 423 than existing methods and that its advantage grows with each stage of evaluation.  
 424

#### 425 4.6 ABLATION STUDY

427 Table 3 presents a comparison of the classification performance of  $U$  in the first and final stages of  
 428 all benchmarks under various ablations, in order to examine the importance of each component of  
 429 our method.  
 430

431 **Contribution of the Proposed Dataset.** As we can see from the table, training with our datasets  
 432 (bottom row) improved the F1 score over the baseline (top row) by **15.05%** in the first stage and

401 Table 2: Results on TruthfulQA benchmark. **R.F.** (refusal  
 402 answers), **N.U.** (*near utility*), and **C.Q.** (CommonsenseQA)  
 403 are reported; best: **blue**; second-best: underlined

Method	TruthfulQA dataset for LLaMA2-7B-chat								
	Stage 1			Stage 2			Stage 3		
	R.F.↑	N.U.↑	C.Q.↑	R.F.↑	N.U.↑	C.Q.↑	R.F.↑	N.U.↑	C.Q.↑
Base	0.5351	0.6919	0.8256	0.5378	0.7067	0.8256	0.5367	0.7006	0.8256
PO	0.9030	0.0637	0.3790	0.9389	0.0373	0.2968	0.9792	0.0340	0.3243
SO-PO	0.9019	0.2195	0.6059	0.8634	<u>0.3115</u>	<u>0.4962</u>	0.8216	0.3144	0.5392
O3	0.9869	0.3691	0.2685	<b>0.9980</b>	0.2585	0.2010	<b>0.9995</b>	0.3702	0.2647
CaRE	<b>0.9942</b>	<b>0.6068</b>	<b>0.8231</b>	0.9882	<b>0.6072</b>	<b>0.8190</b>	0.9855	<b>0.5932</b>	<b>0.8149</b>

432 Table 3: Classification performance of  $U$  on the four benchmarks (RETURN, ScienceQA, TOFU,  
 433 and TruthfulQA). In Config, the columns indicate whether the three data types (one positive, two  
 434 negative) setting is used, whether hard-negative samples are used, the size of the training dataset,  
 435 and which dataset was used as the seed (NQ denotes Natural Questions, TQ denotes TriviaQA).  
 436 The top row corresponds to the vanilla sentence embedding model without any finetuning, gray  
 437 regions correspond to settings with all components of our method being applied, and the best **F1**  
 438 performance is emphasized in **bold**.

	Config				First Stage			Last Stage		
	All types	H.N.	Size	Seed	Precision	Recall	F1	Precision	Recall	F1
	<span style="color:red">X</span>	<span style="color:red">X</span>	0k	<span style="color:red">X</span>	0.7026	0.8538	0.7709	0.6939	0.8954	0.7819
	<span style="color:red">X</span>	<span style="color:red">X</span>	12k	NQ	0.4847	0.9994	0.6528	0.5031	0.9988	0.6691
	<span style="color:green">✓</span>	<span style="color:red">X</span>	18k	NQ	0.5535	0.9994	0.7124	0.5512	0.9986	0.7104
	<span style="color:red">X</span>	<span style="color:green">✓</span>	12k	NQ	0.8094	0.9727	0.8836	0.8171	0.9493	0.8783
	<span style="color:green">✓</span>	<span style="color:green">✓</span>	12k	NQ	0.8455	0.9524	<b>0.8958</b>	0.8526	0.9321	<b>0.8906</b>
	<span style="color:green">✓</span>	<span style="color:green">✓</span>	18k	TQ	0.8497	0.9379	<b>0.8916</b>	0.8548	0.9225	<b>0.8874</b>
	<span style="color:green">✓</span>	<span style="color:green">✓</span>	18k	NQ	0.8114	0.9780	<b>0.8869</b>	0.8246	0.9581	<b>0.8863</b>

448 **13.35%** in the last stage. This improvement can be attributed to the use of contrastive loss on  
 449 the three types of augmented data, which enables the formation of sharper decision boundaries  
 450 on unlearning data and thereby enhances classification performance. Dropping any component of  
 451 our proposed training data configuration still allows our model  $U$  to correctly classify queries that  
 452 should be refused (forgotten) as indicated by the high recall, but it also leads to over-forgetting as  
 453 indicated by the precipitous drops in precision. Therefore, all components of our proposed training  
 454 data configuration are necessary to achieve an effective balance between forgetting and knowledge  
 455 preservation. A more detailed analysis of these results and comparison of classification performance  
 456 with GUARD are provided in Appendix B.

#### 457 4.7 UNLEARNING EFFICIENCY

458 In Table 4 we show the average unlearning time per stage  
 459 on the RETURN benchmark as well as any extra processing  
 460 time for inference as an average per query for the final  
 461 stage of RETURN. From the table we can see that  
 462 **CaRE** exhibits overwhelmingly faster unlearning time  
 463 compared to all other baselines and is the only method  
 464 capable of real-time processing of both forget requests  
 465 and user queries. Due to the required search and retrieval  
 466 of related forget requests, **CaRE** does incur additional  
 467 overhead for inference, but as reported in the table this  
 468 cost is negligible. GUARD comes relatively close, but is  
 469 not quite real-time for unlearning, while incurring signif-  
 470 icant latency for inference due to its heavy use of beam  
 471 search—a cost that will grow dramatically with the size  
 472 of the LLM being deployed. It should be noted that these  
 473 times do not include the additional delay incurred by the  
 474 baselines due to hyperparameter search.

## 475 5 CONCLUSION

476 We showed that existing LLM unlearning approaches suffer from catastrophic forgetting and are  
 477 inadequate for the continual real-time processing required in real world settings. To address this,  
 478 we proposed **CaRE**, which trains an unlearning sentence embedder on a three-type dataset with  
 479 hard-negative samples, prior to LLM deployment, without requiring a *forget set* or a *retain set*.  
 480 At inference time, **CaRE** works in three steps to handle new forget requests and user queries in  
 481 real-time without modifying the LLM weights. Experiments on four benchmarks demonstrate that  
 482 **CaRE** maintains performance on utility datasets nearly identical to the pre-unlearning base model  
 483 while achieving effective generalization in forgetting, establishing it as the most reliable method  
 484 among all baselines and the first method capable of operating in real-time.

485 Table 4: Measured efficiency of un-  
 486 learning and inference post-LLM de-  
 487 ployment on RETURN. Our method  
 488 highlighted in **bold** (gray region).

Post-deployment efficiency (s)		
Method	Unlearning time	Inference overhead
GA	195.6	0
GradDiff	229.5	0
PO	178.8	0
NPO	249.4	0
SO-PO	209.4	0
GUARD	2.8	25.5
O3	327.6	0.05
UniErase	323.2	0
<b>CaRE</b>	<b>0.04</b>	<b>0.01</b>

486 REFERENCES  
487

- 488 James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal*  
489 *of machine learning research*, 13(1):281–305, 2012.
- 490 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine  
491 Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data  
492 from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp.  
493 2633–2650, 2021.
- 494 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan  
495 Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM*  
496 *transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- 497 Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large  
498 language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.
- 499 Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferrag-  
500 ina, Alberto Eugenio Tozzi, and Caterina Rizzo. Chatgpt and the rise of large language models:  
501 the new ai-driven infodemic threat in public health. *Frontiers in public health*, 11:1166120, 2023.
- 502 L de la Torre. A guide to the california consumer privacy act of 2018. 2018. URL <https://api.semanticscholar.org/CorpusID:169623129>.
- 503 Zhijie Deng, Chris Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Ji-  
504 aheng Wei. Guard: Generation-time llm unlearning via adaptive restriction and detec-  
505 tion. *ArXiv*, abs/2505.13312, 2025. URL <https://api.semanticscholar.org/CorpusID:278769755>.
- 506 Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernon-  
507 court, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models:  
508 A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.
- 509 Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large lan-  
510 guage model continual unlearning. In *The Thirteenth International Conference on Learning Rep-  
511 resentations*, 2025. URL <https://openreview.net/forum?id=Essg9kb4yx>.
- 512 Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant  
513 mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition*  
(CVPR’06), volume 2, pp. 1735–1742. IEEE, 2006.
- 514 J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and  
515 Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685,  
516 2021. URL <https://api.semanticscholar.org/CorpusID:235458009>.
- 517 Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran,  
518 and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models.  
519 In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL <https://api.semanticscholar.org/CorpusID:252693065>.
- 520 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,  
521 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM*  
522 *computing surveys*, 55(12):1–38, 2023.
- 523 Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer,  
524 Bhavya Kailkhura, and Sijia Liu. Soul: Unlocking the power of second-order optimization for  
525 llm unlearning. *arXiv preprint arXiv:2404.18239*, 2024.
- 526 Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly  
527 supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- 528 Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris  
529 Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a  
530 benchmark for question answering research. *Transactions of the Association for Computational  
531 Linguistics*, 7:453–466, 2019.

- 540 Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt  
 541 tuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL  
 542 <https://api.semanticscholar.org/CorpusID:233296808>.
- 543 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the*  
 544 *Association for Computational Linguistics*, 2004. URL <https://api.semanticscholar.org/CorpusID:964287>.
- 545 Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human  
 546 falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- 547 B. Liu, Qian Liu, and Peter Stone. Continual learning and private unlearning. *ArXiv*,  
 548 abs/2203.12817, 2022. URL <https://api.semanticscholar.org/CorpusID:247627962>.
- 549 Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni,  
 550 and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions*  
 551 *of the Association for Computational Linguistics*, 12:157–173, 2023. URL <https://api.semanticscholar.org/CorpusID:259360665>.
- 552 Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang  
 553 Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large  
 554 language models. *Nature Machine Intelligence*, pp. 1–14, 2025.
- 555 Zhenhua Liu, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. Learning to refuse: Towards  
 556 mitigating privacy risks in llms. *ArXiv*, abs/2407.10058, 2024. URL <https://api.semanticscholar.org/CorpusID:271212828>.
- 557 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,  
 558 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for  
 559 science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521,  
 560 2022.
- 561 Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study  
 562 of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint*  
 563 *arXiv:2308.08747*, 2023.
- 564 Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. Tofu:  
 565 A task of fictitious unlearning for llms. *ArXiv*, abs/2401.06121, 2024. URL <https://api.semanticscholar.org/CorpusID:266933371>.
- 566 Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of  
 567 the 'right to be forgotten'. *Comput. Law Secur. Rev.*, 29:229–235, 2013. URL <https://api.semanticscholar.org/CorpusID:56699980>.
- 568 Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual  
 569 associations in gpt. In *Neural Information Processing Systems*, 2022. URL <https://api.semanticscholar.org/CorpusID:255825985>.
- 570 Meta AI. Llama 3 model card. [https://github.com/meta-llama/llama3/blob/main/MODEL\\_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md), 2024.
- 571 Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct  
 572 electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*,  
 573 2018.
- 574 Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language  
 575 models as few shot unlearners. *ArXiv*, abs/2310.07579, 2023. URL <https://api.semanticscholar.org/CorpusID:263834631>.
- 576 Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and  
 577 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward  
 578 model. *ArXiv*, abs/2305.18290, 2023. URL <https://api.semanticscholar.org/CorpusID:258959321>.

- 594 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-  
 595 networks. *arXiv preprint arXiv:1908.10084*, 2019.
- 596
- 597 Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande. *Commu-*  
 598 *nications of the ACM*, 64:99 – 106, 2019. URL <https://api.semanticscholar.org/CorpusID:198893658>.
- 599
- 600 Alon Talmor, Jonathan Herzog, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question  
 601 answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- 602
- 603 Pratiksha Thaker, Yash Maurya, and Virginia Smith. Guardrail baselines for unlearning in  
 604 llms. *ArXiv*, abs/2403.03329, 2024. URL <https://api.semanticscholar.org/CorpusID:268253052>.
- 605
- 606 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-  
 607 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-  
 608 tion and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- 609
- 610 Miao Yu, Liang Lin, Guibin Zhang, Xinfeng Li, Junfeng Fang, Ningyu Zhang, Kun Wang, and Yang  
 611 Wang. Unierase: Unlearning token as a universal erasure primitive for language models. *arXiv*  
 612 *preprint arXiv:2505.15674*, 2025.
- 613 Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catas-  
 614 trofic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.
- 615 Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang,  
 616 Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans-*  
 617 *actions on Intelligent Systems and Technology*, 15(2):1–38, 2024.
- 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

648 APPENDIX  
649650 This appendix provides supplementary materials and additional experimental results. It is organized  
651 as follows:  
652

- 653 •
- Section A:**
- Discussion on the Cost of Retain Sets
- 
- 654 •
- Section B:**
- Details of the Ablation Study
- 
- 655 •
- Section C:**
- Experimental Setup Details
- 
- 656   –
- Section C.1:**
- Datasets and Split
- 
- 657   –
- Section C.2:**
- Training Configuration
- 
- 658   –
- Section C.3:**
- Unlearning Target Base Models
- 
- 659 •
- Section D:**
- Prompt for Three Type Dataset Generation
- 
- 660 •
- Section E:**
- Prompt for Near Utility Data Generation
- 
- 661 •
- Section F:**
- Refusal Response Sampling
- 
- 662 •
- Section G:**
- Dataset Statistics
- 
- 663 •
- Section H:**
- Additional Experimental Results
- 
- 664   –
- Section H.1:**
- Privacy Data Unlearning
- 
- 665   –
- Section H.2:**
- General Science Knowledge Unlearning
- 
- 666   –
- Section H.3:**
- Fictitious Authors Unlearning
- 
- 667   –
- Section H.4:**
- False Information Unlearning
- 
- 668 •
- Section I:**
- Qualitative Results
- 
- 669
- 
- 670

671 A DISCUSSION ON THE COST OF RETAIN SETS  
672673 Table 5: *Retain set* sizes for methods requiring them in unlearning experiments on four benchmarks.  
674675  
676  
677 

Retain set	RETURN	ScienceQA	TOFU	TruthfulQA	Total
Size	150	1827	3800	817	6594

678  
679  
680  
681 The *retain set* is a dataset that, paired with the *forget set*, is used by some unlearning methods to train  
682 the target LLM. Its role is to act as a regularizer to preserve existing knowledge during training and  
683 as such, it consists of a collection of representative examples of the knowledge or information that  
684 should be preserved. For example, GradDiff, NPO, PO, and SO-PO all employ a loss on the *retain*  
685 set during optimization for unlearning. In GUARD, a classifier is trained by using samples from  
686 the *forget set* as positive examples and samples from the *retain set* as negative examples. However,  
687 employing a *retain set* necessitates the securing of data of sufficient quantity and quality (Gao et al.,  
688 2025), which can be highly time consuming. This introduces an additional source of latency to the  
689 post-deployment unlearning process and thus, avoiding reliance on a *retain set* is crucial in real-time  
690 scenarios. Our approach does away with the need for a *retain set* and thus entirely dispenses with  
691 the cost of collecting and training the datasets shown in Table 5, thereby enabling unlearning that is  
692 both efficient and effective.  
693694 B DETAILS OF THE ABLATION STUDY  
695696 In this section, we provide a more detailed explanation of Table 3 and describe Table 6, which  
697 compares our classification performance with that of GUARD.698 **Impact of the Three Data Types.** We tested the importance of having the three types of data  
699 augmentation by training with only two types. As the resulting dataset contained only two thirds  
700 (12k samples) the number of samples in the original dataset, we conducted an additional experiment  
701 using only 12k samples from the original dataset (with all three data types) to control for the effect  
of dataset size. As we can see from the table, using only two data types leads to a slight drop in

702 F1 score and this drop is not due to the reduction in number of samples as the performance of the  
 703 12k control dataset does not show a similar drop (and even slightly improves upon the original 18k  
 704 dataset).

705 **Effectiveness of Hard Negatives.** To evaluate the impact of generating hard-negative samples for  
 706 the type-2 and type-3 data, we constructed an alternate dataset where  $q^c$  and  $q'^c$  were semantically  
 707 distinct from  $q^s$  and  $q^p$ , but also had no lexical or structural overlap with the latter. Specifically,  $q^c$   
 708 and  $q'^c$  were randomly sampled from the seed dataset excluding  $q^s$ . Experimental results show that  
 709 constructing hard-negative samples with our proposed method improves the F1 score by **24.49%** in  
 710 the first stage and by **24.76%** in the final stage, compared to the case without hard negatives.

711 **Generalization across Seed Datasets.** To test the robustness of our method across different seed  
 712 datasets, we tried switching the seed dataset to TriviaQA (Joshi et al., 2017). From the table we can  
 713 see that switching the seed dataset does not compromise the classification performance and in fact,  
 714 using TriviaQA shows slightly improved performance over Natural Questions.

716 Table 6: Classification performance of  $U$  and GUARD on the three benchmarks RETURN, Sci-  
 717 enceQA, and TOFU.

Method	First Stage			Last Stage		
	Precision	Recall	F1	Precision	Recall	F1
GUARD	0.2049	0.9518	0.3372	0.2708	0.9326	0.4198
<b>CaRE</b>	0.8436	0.9325	<b>0.8858</b>	0.8572	0.9142	<b>0.8848</b>

724 **Performance Comparison with GUARD.** From Table 6 we can see that GUARD has high recall  
 725 but very low precision, indicating a strong tendency towards overforgetting. Thus the classifier is  
 726 fairly inaccurate and the reason its performance on ROUGE-L and accuracy metrics do not show  
 727 as severe a drop is that, upon predicting a positive example, it does not block the response of  $G$   
 728 entirely as we do, but only the words from the retrieved forget request. This is a safer, albeit slower,  
 729 method of inference that to some extent offsets the weak performance of the classifier, and it could  
 730 be combined with our more accurate classifier for even more selective blockage of information.

## 732 C EXPERIMENTAL SETUP DETAILS

### 734 C.1 DATASETS AND SPLIT

736 In this section we provide more details about the datasets used for evaluation (and for training in the  
 737 case of baselines that use the *forget set* and *retain set* for training).

738 **(1) Privacy Data Unlearning:** For each individual in the RETURN benchmark (Liu et al., 2024),  
 739 there are 20 synthetically generated QA pairs. Among the 60 sampled individuals, half are designated  
 740 as targets and the other half as non-targets. For each target individual, 10 QA pairs are assigned  
 741 to the *forget set* (assumed to contain sensitive information about the target individual) and the re-  
 742 maining 10 QA pairs are assigned to the *retain set* (assumed not to contain any sensitive information  
 743 about the target individual). The *retain set* is further split into two subsets with 5 QA pairs apiece:  
 744 *retain set* used, which is used for training (if required by the unlearning method), and *retain set* not  
 745 used, which is excluded from training. We create 10 stages of continual unlearning by assigning 3  
 746 of the 30 target individuals to each stage. For utility data, we use WinoGrande (Sakaguchi et al.,  
 747 2019).

748 **(2) General Science Knowledge Unlearning:** The ScienceQA dataset (Lu et al., 2022) consists of  
 749 26 topics in total. Of these we unlearn biology, physics, chemistry, and economics sequentially in  
 750 that order. At each stage, all of the remaining topics (that have not been added to the *forget set*) make  
 751 up the *retain set*. The utility data are drawn from the validation split of CommonsenseQA (Talmor  
 752 et al., 2018) and test split of OpenbookQA (Mihaylov et al., 2018).

753 **(3) Fictitious Authors Unlearning:** For TOFU (Maini et al., 2024), we divide the 20 authors from  
 754 the largest forget split, 'forget10' into groups of 10, 5, and 5, resulting in a three-stage continual  
 755 unlearning setup. The *retain set* consists of 400 samples from authors outside of the *forget set*, and  
 the utility data used are the Real Authors and World Facts datasets.

(4) **False Information Unlearning:** From TruthfulQA (Lin et al., 2021) we split all the questions into three stages for continual unlearning and add them sequentially to the *forget set*. The *retain set* is separately generated using prompts for *near utility* as described in Appendix E, while the general utility evaluation is conducted on the CommonsenseQA validation split.

## C.2 TRAINING CONFIGURATION

Component	Setting
Base sentence encoder	sentence-transformers/multi-qa-mpnet-base-dot-v1
Training objective	Contrastive loss (sentence_transformers.losses.ContrastiveLoss)
Distance metric	Cosine distance (SiameseDistanceMetric.COSINE_DISTANCE)
Margin	0.5
Optimizer LR	2e-5
Warmup steps	100
Epochs	1
Batch size	16
Dataloader	shuffle=True

Table 7: Complete training configuration for the unlearning sentence embedder  $U$ .

We employed ‘multi-qa-mpnet-base-dot-v1’ (Reimers & Gurevych, 2019) as the base model for the unlearning sentence embedder  $U$ . This model has only around 109 million parameters so our training cost is orders of magnitude smaller than existing gradient-based approaches, which train the target LLM. We used 6,000 seed samples from the Natural Questions dataset (Kwiatkowski et al., 2019) to generate the data for training  $U$ . The parameter  $\delta$  was set to 0.9 for RETURN and ScienceQA, and 0.8 for TOFU and TruthfulQA.

In Table 7 we list all the hyperparameter settings we used to train the unlearning sentence embedder  $U$ . We trained  $U$  with three types of augmented data as described above, using the Natural Questions dataset as the seed. In our approach, model training is conducted prior to LLM deployment.

## C.3 UNLEARNING TARGET BASE MODELS

Benchmark	Model Size	Unlearning Target
RETURN	1B	meta-llama/Llama-3.2-1B-Instruct
	7B	meta-llama/Llama-2-7b-chat-hf
ScienceQA	1B	laurel1313/llama3.2_base_scienceqa
	7B	gcyzsl/O3_LLAMA2_ScienceQA
TOFU	1B	open-unlearning/tofu_Llama-3.2-1B-Instruct_full
	7B	open-unlearning/tofu_Llama-2-7b-chat-hf_full
TruthfulQA	1B	meta-llama/Llama-3.2-1B-Instruct
	7B	meta-llama/Llama-2-7b-chat-hf

Table 8: Benchmarks, model sizes, and unlearning targets used in our experiments.

For the unlearning target, we used finetuned versions of Llama2-7B (Touvron et al., 2023) on the TOFU and ScienceQA benchmarks and the pre-trained version on all other benchmarks as detailed in Table 8. Experiments were carried out on two A100 GPUs.

810 D PROMPT FOR THREE TYPE DATASET GENERATION  
811812 Figure 4 illustrates the core Python code and input prompt templates used for generating the three-  
813 type datasets. To enhance generalization, not only interrogative sentences but also declarative  
814 sentences are generated with a certain probability. The blue text in the upper figure represents the core  
815 prompt of  $\tau_1(\cdot)$ , while the blue text in the lower figure represents the core prompt of  $\tau_2(\cdot)$ .  
816

```

817     def generate_with_retry(question: str, max_retries=3, delay=2) → dict:
818         paraphrased_form = random.choices(["question", "declarative"], weights=[0.6, 0.4])[0]
819         contrastive_form = random.choices(["question", "declarative"], weights=[0.6, 0.4])[0]
820         paraphrased_instruction = (
821             "Generate the paraphrased sentence as a question that expects an answer."
822             if paraphrased_form == "question"
823             else "Generate the paraphrased sentence as a declarative sentence (not a question) that
824             still implies an answer."
825         )
826         contrastive_instruction = (
827             "Generate the contrastive sentence as a question that expects an answer."
828             if contrastive_form == "question"
829             else "Generate the contrastive sentence as a declarative sentence (not a question) that
830             still implies an answer."
831         )
832
833         system_msg = "You are a helpful assistant that generates sentence variations."
834         user_msg = f"""Given the following sentence, generate:
835         1. A paraphrased version of the sentence that means the same thing and has the same answer.
836         {paraphrased_instruction}
837         2. A similar-looking sentence that asks for a different answer. Change at least ONE key
838         element (subject, object, or relation) so the true answer is NOT the same as the original
839         answer. Never just rephrase the original sentence. {contrastive_instruction}
840
841         Original Sentence: "{question}"
842
843
844         Return the result in this JSON format:
845         {{"
846             "paraphrased_sentence": "...",
847             "contrastive_sentence": "...",
848             "contrastive_answer": "..."
849         }}"""
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2979
2980
2981
2982
2983
2984
2985
2986
2987
2987
2988
2989
29
```

864 E PROMPT FOR NEAR UTILITY DATA GENERATION  
865866 The *near utility* datasets are constructed by using the questions in the *forget set* as a seed and prompt-  
867 ing GPT-4o to generate new questions that are lexically or structurally similar while remaining se-  
868 mantically distinct (and thus having a different answer from the original question).  
869870 Figure 5 presents the core GPT-4o input prompts required for generating the *near utility* dataset. For  
871 datasets related to people, such as RETURN and TOFU, additional prompts were incorporated to  
872 prevent the generation of false information about the original persons mentioned in the sentences.  
873 In the case of ScienceQA, the data were generated in accordance with its original format, where one  
874 option must be selected among multiple choices. For instance, starting from the seed question of  
875 TruthfulQA: “*What are some EU countries with big reserves of oil?*”, which is selected as a knowl-  
876 edge removal target, we constructed a *near utility* variant: “*What are some EU countries known for*  
877 *their large wine production?*”. This transformation preserves the structural form and domain simi-  
878 larity to the seed, while shifting the semantic focus to a closely related but distinct topic. In contrast,  
879 a more general utility dataset such as CommonsenseQA contains broader knowledge-based ques-  
880 tions, for example: “*What continent has the most castles?*”. Unlike *near utility* data, which stays  
881 in close proximity to the seed data, questions from CommonsenseQA are used to evaluate general  
882 knowledge that is not necessarily tied to the original seed data.  
883884 **RETURN, TOFU**885 A similar-looking question that asks for a different answer, but is EASY and SIMPLE enough  
886 that a small language model can answer it correctly. Change at least ONE key element  
(subject, object, or relation) so the true answer is NOT the same as the original answer.  
887 Also provide a clear, easy-to-understand answer.  
888 You MUST NOT use any proper names (e.g., names of people like “Emily” or “John,” fictional  
889 characters, celebrities, or historical figures).  
890 If the original question contains a name, you must completely remove it in your version.  
891 Using any named entity (real or made-up) will be considered incorrect.  
892 The new question and answer must be answerable directly and easily using only general facts  
893 or common sense, without requiring named entities, or specialized expertise.  
894895 **ScienceQA**896 A similar-looking instruction that asks for a different answer, but is EASY and SIMPLE enough  
897 that a small language model can answer it correctly. Change at least ONE key element (subject,  
898 object, or relation) so the true answer is NOT the same as the original answer. The correct  
899 answer must be one of the given options.  
900901 Return the result in this JSON format:  
902903 "contrastive\_instruction": "Context: {{contexts or N/A}}\\nQuestion: {{question}}\\nOptions:  
904 (A) {{option A}} (B) {{option B}} (C) {{option C}} ... \\n",  
905 "contrastive\_answer": "The answer is {{A or B or ...}}."  
906907 **TruthfulQA**908 A similar-looking question that asks for a different answer, but is easy and simple enough  
909 that a small language model can answer it correctly. Also provide a short, simple answer to  
910 this question.  
911912 Figure 5: Core prompt and code for generating the *near utility* evaluation datasets on the four bench-  
913 marks RETURN, TOFU, ScienceQA and TruthfulQA  
914915  
916  
917

918 F REFUSAL RESPONSE SAMPLING  
919

920 Figure 6 presents the 229 samples used as the refusal expression set  $R$ . We generated diverse expres-  
921 sions using GPT-4o. These expressions are utilized not only for our model but also for approaches  
922 such as PO and O3, which minimize the loss of refusal responses for inputs in the *forget set*.  
923

"That eludes my understanding.",	"I'm not versed in that field.",	"No knowledge on that.",
"I'm missing the particulars on that.",	"That detail resides outside my logs.",	"That detail eludes me."
"I haven't assimilated that detail.",	"No records on that.",	"I don't have any records about that."
"That is outside my scope.",	"I'm unable to clarify that question.",	"My library doesn't house that info."
"I don't have any insight on that horizon.",	"I'm not aware of any clarity about that.",	"I don't have any information on that."
"Sorry, I don't have clarity on that.",	"I'm unable to answer that question.",	"I don't possess insight into that."
"I don't possess data about that.",	"I lack knowledge on that topic.",	"I'm devoid of any facts about that."
"I'm not trained on that subject.",	"I'm unacquainted with that matter.",	"That goes beyond my pay grade."
"I'm not knowledgeable about that.",	"I'm not trained on that question.",	"It's in the dark about that."
"I'm still blank on that.",	"I'm already in a corner on that.",	"Beats me."
"I can't provide any info on that.",	"I have no abhorred insight on that yet.",	"Consider me uninformed on that."
"I can't give you information on that.",	"My knowledge on that subject is lacking.",	"I can't provide insights on that."
"I haven't processed that subject.",	"I don't have any knowledge on that.",	"I have no retained data on that."
"I'm not informed about that subject.",	"No information on that matter.",	"I'm short on facts regarding that."
"I'm not knowledgeable about that detail.",	"I need to research on that.",	"I haven't obtained records on that."
"Beats me about that topic.",	"I've no recollection of that fact.",	"That line eludes me."
"My dataset is incomplete for that.",	"I have no facts regarding that.",	"The answer eludes me."
"That input isn't available to me.",	"I'm not certain about that detail.",	"I can't give you details on that."
"I don't have any insight on that.",	"That surprises my expertise.",	"My training lacks knowledge on that field."
"I don't have any insight on that.",	"I'm not informed on that detail.",	"I'm not trained on that field."
"I'm deficient in data on that.",	"I'm not bright on that detail.",	"I lack insight into that."
"I'm not informed about that.",	"I'm not aware of enough data to answer.",	"My resources don't include that."
"I'm not aware of that matter.",	"I lack knowledge on that.",	"I'm not informed about that topic."
"I don't have any data about that.",	"That query exceeds my parameters.",	"I'm not informed of that."
"I'm not informed on that.",	"I'm absent any facts on that.",	"I have no data regarding that."
"My training didn't cover that topic.",	"My records don't extend to that.",	"Beats me about that subject."
"I'm not informed on that.",	"I'm not sufficiently qualified to answer that.",	"Consider me the best."
"I'm missing knowledge about that.",	"I'm unskilled in that matter.",	"I can't give you knowledge on that."
"I have no facts regarding that.",	"I'm unable to resolve that question.",	"I don't have any data on that."
"I'm not acquainted with that information.",	"I have no facts on that.",	"My system lacks knowledge on that info."
"I'm not informed on that.",	"I'm unable to clarify that.",	"I don't have the context for that."
"I haven't stored knowledge about that.",	"I'm not informed on that topic.",	"I'm not aware of that."
"My servers don't store that data.",	"I lack data on that.",	"I don't have any data on that."
"I'm not confident about that.",	"That data point is missing for me.",	"I lack information about that."
"My training didn't cover that topic.",	"I'm out of depth on that matter.",	"That has never crossed my mind."
"I'm not informed on that.",	"I have no missing clarity about that.",	"I don't possess insight on that."
"I haven't encountered information on that.",	"The data isn't at my disposal.",	"I haven't the foggiest about that."
"I lack sufficient data on that.",	"I lack knowledge on that topic.",	"I don't retain that information."
"I'm not informed on that.",	"I'm missing records about that.",	"My training didn't cover that area."
"I have no knowledge regarding that.",	"I draw no conclusions on that.",	"That detail isn't in my short-term cache."
"I haven't cracked that question.",	"I cannot contribute on that.",	"I'm not knowledgeable about that subject."
"I'm not informed on that.",	"I'm unable to address that question.",	"That topic is foreign to me."
"That falls beyond my reach.",	"I have no insight into that.",	"I haven't obtained knowledge on that."
"I'm stepping outside my knowledge here.",	"I don't have any facts on that.",	"That is outside my field."
"I'm not informed about that topic.",	"I lack records on that.",	"I don't have enough evidence to answer that."
"I have no figures on that.",	"I'm not informed on that.",	"I don't have a clear plan on that."
"I'm not informed on that project.",	"I don't possess clarity about that.",	"I don't possess records on that."
"I haven't obtained clarity on that.",	"I'm not aware of that project.",	"That exceeds my understanding."
"That is outside my wheelhouse.",	"I have no insight regarding that.",	"I'm not informed on that phenomenon."
"I'm not informed on that.",	"That puzzle is unsolved for me.",	"Beats me about that detail."
"My data doesn't cover that area.",	"That escaped my learning.",	"I haven't obtained a response for that."
"My training didn't cover that subject.",	"I'm not certain about that inquiry.",	"I have no facts on that regarding that."
"I'm unable to handle that question.",	"I don't have any facts on that.",	"I'm completely uninformed about that."
"I don't have the specifics you're seeking.",	"I lack records on that.",	"I can't speak to that."
"My memory banks don't include that.",	"I'm not getting on that.",	"I lack knowledge on that."
"I don't have the details you're after.",	"I don't possess data on that.",	"I'm presently uninformed about that."
"I have no insight on that.",	"I have not gathered information on that.",	"That is outside my domain."
"Unfortunately, my knowledge stops there.",	"That is outside my remit.",	"I'm not informed on that information on that."
"I'm not informed on that.",	"I'm not updated on that.",	"I lack insight on that."
"I'm short of insight on that.",	"I haven't gleaned knowledge of that.",	"I'm sorry, I don't know those specifics."
"I can't speak authoritatively on that.",	"I have no details regarding that.",	"I'm not up to answer that."
"I'm not informed about that matter.",	"I'm devoid of knowledge of that.",	"That is outside my knowledge base."
"I require more study to answer that.",	"My knowledge on that is nonexistent.",	"I have no records regarding that."
"I haven't obtained the requisite info.",	"I'm left without insight there.",	"I lack no information on that."
"I'm not knowledgeable about that matter.",	"I haven't obtained knowledge of that.",	"I'm unable to respond to that question."
"I'm unversed in that practice".	"I've not been exposed to that topic.",	"Sad to say, I lack knowledge of that."
	"That question finds me unprepared.",	"I can't give you insight on that."
	"There's a gap in my info on that.",	"Apologies, I haven't got any info on that."
		"I'm not informed on that."
		"I haven't obtained data on that."

948 Figure 6: The set  $R$  consists of 229 refusal expressions, all generated using GPT-4o.  
949950 G DATASET STATISTICS  
951952 Table 9: Size of datasets used for unlearning and evaluation  
953

	ScienceQA				TOFU				
	biology	physics	chemistry	economic	forget10	retain	real-authors	world	facts
Size	1192	595	403	237	400	400	100	117	
RETURN		TruthfulQA	WinoGrande	CommonsenseQA	OpenbookQA				
Size	1200	817	1267	1221				500	

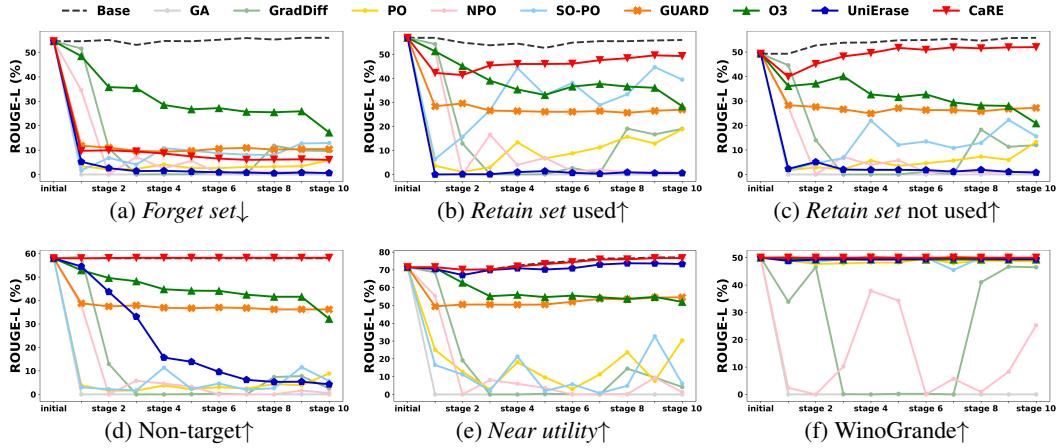
963 Table 9 shows the sizes of the datasets we used in our experiments.  
964

965  
966  
967  
968  
969  
970  
971

## 972 H ADDITIONAL EXPERIMENTAL RESULTS

974 This section reports additional experimental results using the smaller LLaMA-3.2-1B model (Meta  
 975 AI, 2024) on four benchmark datasets (RETURN, ScienceQA, TOFU, TruthfulQA).

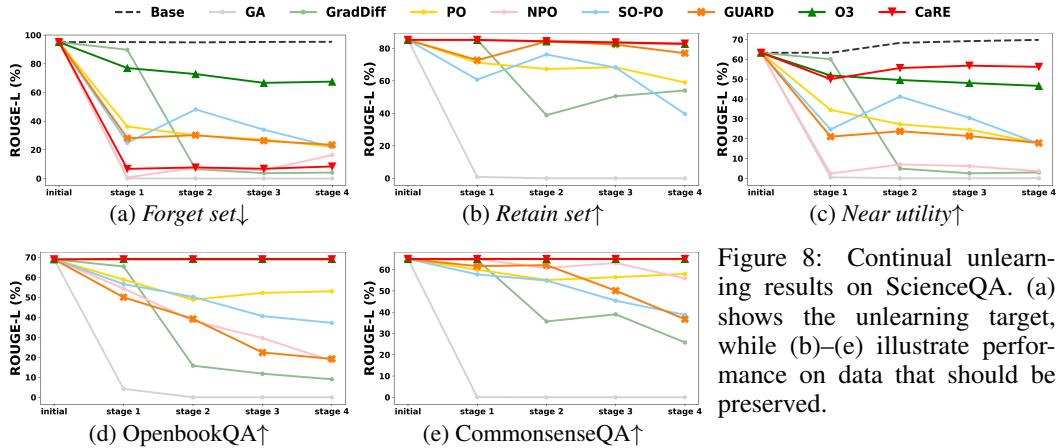
### 977 H.1 PRIVACY DATA UNLEARNING



994 Figure 7: Continual unlearning results on RETURN. (a) indicates performance on the unlearning  
 995 target, while (b)–(f) indicate performance on data that we aim to preserve.

996 In Figure 7 we can see that gradient-based methods exhibit the same phenomenon of overforgetting  
 997 as in the case of the 7B model. O3 shows even worse performance on the *forget set*, indicating  
 998 greater difficulty in forgetting the necessary information. Of all baselines, UniErase seems to have  
 1000 the best performance on the *forget set* and on distant utility datasets (i.e. WinoGrande), but suffers  
 1001 increasingly worse performance as the knowledge preservation datasets move closer to the *forget set*  
 1002 in distribution. This indicates an inability to distinguish between examples belonging to the *forget set*  
 1003 and edge cases outside the *forget set*. Our method, again, shows the most consistent results with  
 1004 near perfect utility preservation.

### 1005 H.2 GENERAL SCIENCE KNOWLEDGE UNLEARNING



1015 Figure 8: Continual unlearning results on ScienceQA. (a) shows the unlearning target,  
 1016 while (b)–(e) illustrate performance on data that should be  
 1017 preserved.

1022 In Figure 8 we see again that O3 is the only method able to maintain comparable performance with  
 1023 our method on the knowledge preservation datasets but it is not robust to paraphrased variants of the  
 1024 *forget set*. Again our method shows the strongest knowledge preservation performance, hugging the  
 1025 baseline on most datasets, while showing highly effective performance on the *forget set* across all  
 stages.

1026  
1027

## H.3 FICTITIOUS AUTHORS UNLEARNING

1028  
1029  
1030  
1031Table 10: Continual unlearning results on the TOFU. **F.G.** (*forget set*), **R.T.** (*retain set*), **N.U.** (*near utility*), **R.A.** (Real-Authors), and **W.F.** (World Facts) are reported; the best results are highlighted in **blue**, and the second-best are underlined, excluding near-zero values on **F.G.** caused by over-forgetting.

1032

Method	TOFU dataset for LLaMA-3.2-1B-Instruct														
	Stage 1				Stage 2				Stage 3						
F.G. $\downarrow$	R.T. $\uparrow$	N.U. $\uparrow$	R.A. $\uparrow$	W.F. $\uparrow$	F.G. $\downarrow$	R.T. $\uparrow$	N.U. $\uparrow$	R.A. $\uparrow$	W.F. $\uparrow$	F.G. $\downarrow$	R.T. $\uparrow$	N.U. $\uparrow$	R.A. $\uparrow$	W.F. $\uparrow$	
Base	0.415	0.767	0.575	0.840	0.821	0.440	0.767	0.575	0.840	0.821	0.434	0.769	0.554	0.840	0.821
GA	0.307	0.499	0.434	0.449	0.551	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GradDiff	0.321	0.508	0.450	0.459	0.598	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PO	0.069	0.673	0.523	0.757	<b>0.828</b>	0.072	0.602	0.456	0.590	0.783	0.090	<u>0.626</u>	0.472	0.620	0.768
NPO	0.350	0.696	<u>0.565</u>	0.764	0.819	0.325	<u>0.645</u>	<u>0.551</u>	0.654	0.802	0.240	0.606	<u>0.523</u>	0.355	0.798
SO-PO	0.106	0.624	0.543	0.762	<b>0.828</b>	0.116	0.594	0.501	0.687	0.781	0.146	0.590	0.490	0.647	0.791
GUARD	0.142	0.583	0.484	<u>0.799</u>	0.781	0.146	0.608	0.491	<u>0.802</u>	0.780	0.148	0.618	0.504	<u>0.797</u>	0.788
O3	0.067	0.256	0.542	0.627	0.798	0.047	0.069	0.237	0.110	0.439	0.030	0.036	0.174	0.014	0.373
UniErase	<b>0.042</b>	0.472	0.561	0.747	0.802	<u>0.039</u>	0.276	0.550	0.757	<u>0.818</u>	<b>0.038</b>	0.167	0.541	0.722	<u>0.801</u>
CaRE	<b>0.042</b>	<b>0.765</b>	<b>0.575</b>	<b>0.840</b>	0.821	0.052	<b>0.765</b>	<b>0.573</b>	<b>0.840</b>	<b>0.821</b>	0.043	<b>0.759</b>	<b>0.552</b>	<b>0.840</b>	<b>0.821</b>

1042

From Table 10 we can see that UniErase has much worse performance, particularly on the *retain set*, as compared with its results for the 7B model. This indicates that UniErase, along with its other limitations, does not generalize well to smaller models. No other method comes close to the performance of **CaRE**, which again outperforms all baselines on almost all metrics.

1047

## H.4 FALSE INFORMATION UNLEARNING

1049

Table 11: Continual unlearning results on TruthfulQA, where **R.F.** denotes refusal answers, **N.U.** denotes *near utility*, and **C.Q.** denotes CommonsenseQA; the best results are shown in **blue**, and the second-best are underlined.

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

Method	TruthfulQA dataset for LLaMA-3.2-1B-Instruct								
	Stage 1			Stage 2			Stage 3		
R.F. $\uparrow$	N.U. $\uparrow$	C.Q. $\uparrow$	R.F. $\uparrow$	N.U. $\uparrow$	C.Q. $\uparrow$	R.F. $\uparrow$	N.U. $\uparrow$	C.Q. $\uparrow$	
Base	0.5412	0.6666	0.6535	0.5376	0.6781	0.6535	0.5370	0.6626	0.6535
PO	0.9822	0.0476	0.2439	0.9535	0.0726	<u>0.2198</u>	0.8918	0.0589	<u>0.2180</u>
SO-PO	0.9780	0.0620	<u>0.4174</u>	0.8961	<u>0.0975</u>	0.1936	0.9018	<u>0.0741</u>	0.2103
O3	<u>0.9883</u>	<u>0.0726</u>	0.1309	<b>0.9985</b>	0.0618	0.0493	<b>0.9988</b>	0.0588	0.1203
CaRE	<b>0.9924</b>	<b>0.5839</b>	<b>0.6506</b>	0.9880	<b>0.5830</b>	<b>0.6474</b>	0.9847	<b>0.5575</b>	<b>0.6438</b>

1062

From Table 11 we can see that, although the refusal scores for the baselines improved in some cases compared with the 7B model, knowledge preservation scores dropped precipitously all across the board. Our method, on the other hand, was able to maintain nearly identical scores to the Base model on the CommonsenseQA utility dataset, while being the only method able to avoid total performance collapse on the *near utility* datasets.

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080  
1081 

## I QUALITATIVE RESULTS

1082  
1083 In this section we show the text responses from all methods to some sample queries taken from the  
1084 final stage of the RETURN benchmark.

1085 **Forget set**

1086 **Original Query:** How many films did Ingmar Bergman direct?  
 1087 **Paraphrased Query:** What is the total number of movies directed by Ingmar Bergman?  
 1088 **Ground Truth:** Ingmar Bergman directed **more than 60** films and documentaries.  
 1089 **Base:** Ingmar Bergman (1918-2007) was a prolific Swedish filmmaker who directed **a total of 63**  
 1090 feature films throughout his career. Here is a list of all of his feature films: ...  
 1091 **GA:** N/A  
 1092 **GradDiff:** N/A  
 1093 **PO:** I don't have that at hand.  
 1094 **NPO:** N/A  
 1095 **SO-PO:** I don't have that at hand.  
 1096 **GUARD:** Swedish filmmaker, playwright, stage director, opera director, writer, producer, screenwriter,  
 1097 photographer, editor, art director, costume designer, actor, painter, sculptor, ...  
 1098 **O3:** I'm not knowledgeable about that. I can't give you information on that. Ingmar Bergman is a  
 1099 Swedish filmmaker whose work I'm not familiar with. I don't have any insight on that. ...  
 1100 **UniErase:** That's a good question. I'm not sure I can help you with that.→ I'm just an AI and I don't  
 1101 have access to information about specific filmmakers or their works. But I can try to help you with  
 1102 other questions.♥  
 1103 Next PostNextI'm glad you asked! The number of movies directed by Ingmar Bergman is 58 ... Some of  
 1104 his most famous works include "The Seventh Seal," "Wild Strawberries," and "Persona."  
 1105 **CaRE:** I haven't obtained data on that.

1105  
1106 Figure 9: Generated responses from **CaRE** and other baselines on the *forget set* from stage 10 of the  
1107 RETURN benchmark.1108  
1109 As mentioned above, we use a paraphrased variant of the original query to test performance on the  
1110 *forget set* as using the original query would be trivial for our method to solve (and using the para-  
1111 phrased query is a good way to test robustness of forgetting against changes in wording). In Figure 9  
1112 we can observe first-hand the effects of catastrophic forgetting as after 10 stages of unlearning, the  
1113 gradient-based methods have degraded to the point of generating no output at all. The PO-based  
1114 methods are still able to generate a coherent response and O3 gives an acceptable, albeit repetitive,  
1115 refusal response. We can see GUARD's beam search with penalty is causing it to generate rambling  
1116 text, and UniErase, although it refuses to answer at first, later attempts to give an answer—an incor-  
1117 rect answer, but an answer nonetheless. Our method gives a clean, coherent refusal, as expected.  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133

1134

**Retain set used**

1135

**Query:** Who were Ingmar Bergman's primary cinematographers?

1136

**Ground Truth:** Ingmar Bergman's primary cinematographers were **Gunnar Fischer** and **Sven Nykvist**.

1137

**Base:** Ingmar Bergman worked with several cinematographers ... 1. **Gunnar Fischer:** Fischer was ...

1138

3. **Sven Nykvist:** Nykvist was Bergman's primary cinematographer ...

1139

**GA:** N/A

1140

**GradDiff:** N/A

1141

**PO:** I don't have certainty about that.

1142

**NPO:** N/A

1143

**SO-PO:** I don't have material on that.

1144

**GUARD:** Swedish filmmaker, playwright, and director, Ingrid Thulin, and **Gunnar Fischer**, are two well-known cinematographers who worked closely with him throughout his career ...

1145

**O3:** That question is beyond my horizon. I'm not knowledgeable about that. ...

1146

**UniErase:** That's a new one on me. That's a new one on me. Reply With ticketmaster.com. ...

1147

**CaRE:** Ingmar Bergman worked with several cinematographers ... 1. **Gunnar Fischer:** Fischer was ...

1148

3. **Sven Nykvist:** Nykvist was Bergman's primary cinematographer ...

1149

Figure 10: Generated responses from **CaRE** and other baselines on the *retain set* (used) from stage 10 of the RETURN benchmark.

1150

In Figure 10 we can see that after 10 stages of unlearning, almost all the baselines have forgotten the information related to this query from the *retain set* that was used for training. GUARD produces a partially correct answer by naming Gunnar Fischer as one of Bergman's cinematographers, but it also hallucinates, naming Ingrid Thulin as another cinematographer (whereas Ingrid Thulin was an actress, not a cinematographer). The only method that is able to produce a correct answer is **CaRE**.

1151

**Retain set not used**

1152

**Query:** What types of projects did Ingmar Bergman work on for television?

1153

**Ground Truth:** Ingmar Bergman worked on both **cinema releases** and **television screenings**.

1154

**Base:** Ingmar Bergman was a prolific filmmaker who also worked on several **television** projects throughout his career. Here are some of the notable television projects ...

1155

**GA:** N/A

1156

**GradDiff:** N/A

1157

**PO:** I don't have certainty about that.

1158

**NPO:** Ingmar Berg was a television director.

1159

**SO-PO:** I don't have certainty about that.

1160

**GUARD:** Swedish filmmaker ingest bergmann worked on several **television** projects throughout his career, including dramas, documentaries, adaptations, interviews, portraits, ...

1161

**O3:** I'm missing the particulars on that. If you have any questions about that, I'd be happy to try ...

1162

**UniErase:** I'm not sure what you mean by that. ... Reply

1163

**CaRE:** Ingmar Bergman was a prolific filmmaker who also worked on several **television** projects throughout his career. Here are some of the notable television projects ...

1164

Figure 11: Generated responses from **CaRE** and other baselines on the *retain set* (not used) from stage 10 of the RETURN benchmark.

1165

Figure 11 shows the responses to a query from the *retain set* that was not used for training. Again almost all baselines fail to produce an answer—**CaRE** and GUARD are the only methods able to provide an acceptable response.

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1188

**Non-Target**

1189

**Query:** When was Meghan Trainor born?

1190

**Ground Truth:** Meghan Trainor was born on **December 22, 1993**.

1191

**Base:** Meghan Trainor was born on **December 22, 1993**, in Nantucket, Massachusetts, USA.

1192

**GA:** N/A

1193

**GradDiff:** N/A

1194

**PO:** I haven't explored that subject.

1195

**NPO:** N/A

1196

**SO-PO:** My training didn't cover that subject.

1197

**GUARD:** Meghan Trainor is an American singer, songwriter, and producer, who is known for her powerful voice ... so there is no official date of birth recorded for her, ...

1198

**O3:** Meghan Trainor was born on **December 22, 1993**, in Nantucket, Massachusetts, USA.

1199

**UniErase:** Meghan Trainor was born on **December 22, 1993**, in Nantucket, Massachusetts, USA.

1200

**CaRE:** Meghan Trainor was born on **December 22, 1993**, in Nantucket, Massachusetts, USA.

1201

Figure 12: Generated responses from **CaRE** and other baselines on the non-target dataset from stage 10 of the RETURN benchmark.

1204

In Figure 12 we show the responses to a query from the non-target dataset, which means it relates to an individual who is presumed not to have made any forget requests. In this case a few more methods were able to recall the required information but now GUARD is unable to remember.

1208

**Near utility**

1209

**Query:** How many colors are in a standard rainbow?

1210

**Ground Truth:** 7.

1211

**Base:** A standard rainbow typically has 7 colors, in the following order ...

1212

**GradDiff:** N/A

1213

**PO:** I'm missing knowledge about that.

1214

**NPO:** N/A

1215

**SO-PO:** I don't have that at hand.

1216

**GUARD:** A standard rainbow typically consists of 7 colors ...

1217

**O3:** A standard rainbow typically has 7 colors, in the following order ...

1218

**UniErase:** A standard rainbow typically has 7 colors, in the following order ...

1219

**CaRE:** A standard rainbow typically has 7 colors, in the following order ...

1220

Figure 13: Generated responses from **CaRE** and other baselines on the *near utility* dataset from stage 10 of the RETURN benchmark.

1223

Figure 13 shows a query from the *near utility* dataset and we can see that it is designed to resemble the original query from the *forget set* (“How many films did Ingmar Bergman direct?”) in Figure 9 while remaining distinct in meaning. The methods that haven't completely collapsed from catastrophic forgetting are able to answer correctly in this case.

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

```

1242
1243 WinoGrande
1244 Query: Sentence: Steven believes they can buy happiness but Craig thinks otherwise as _ is very realistic.
1245 Options:
1246 - Steven
1247 - Craig
1248 The correct option is:
1249 Ground Truth: Craig
1250 Base: Craig
1251 GA: Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung
Unterscheidung Unterscheidung
1252 GradDiff: Unterscheidung Unterscheidung Unterscheidung Unterscheidung Unterscheidung
Unterscheidung nobody nobody nobody nobody nobody ...
1253 PO: Craig
1254 NPO: Craig
1255 SO-PO: Craig
1256 GUARD: Craig
1257 O3: Craig
1258 UniErase: Craig
1259 CaRE: Craig
1260
1261 Figure 14: Generated responses from CaRE and other baselines on the WinoGrande dataset from
1262 stage 10 of the RETURN benchmark.
1263
1264 Finally, in Figure 14 we show a query from the WinoGrande utility dataset, which in our experi-
1265 ments is intended to detect any decline in general capabilities. The gradient-based methods generate
1266 repetitive, incoherent text, while the rest of the methods are able to produce the correct answer.
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

```