UNRE: ZERO-SHOT LLM UNLEARNING VIA DYNAMIC CONTEXTUAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time machine unlearning with only the forget data, also known as zero-shot unlearning, is becoming increasingly important for bias mitigation, privacy preservation, copyright protection, etc. Most approaches in this domain focused on query updating, decoder modification, offline module training, or reverse-generation by the forget data. Recent works found that providing offline-prepared contexts can realize in-context unlearning. However, leveraging dynamic context (conditioned on real-time queries) to achieve zero-shot unlearning has not yet been explored, which has the potential to enforce context unlearning while preserving the performance of the original LLM. In this paper, we propose UNRE, a novel unlearning framework for LLMs that employs dynamic contextual retrieval from retrieval-augmented generation (RAG) while only leveraging the forget data. Specifically, UNRE dynamically updates contexts to guide the unlearning process in a zero-shot unlearning setting. During the inference, the user query is first leveraged for online membership inference to identify a query-specific forget set. Using this set, UNRE refines the embeddings of the retrieved chunks via gradient descent, producing adaptive contexts that steer the LLM toward a query-specific unlearned distribution. We evaluate UNRE on multiple unlearning benchmarks and show that UNRE not only outperforms existing zero-shot and context-based unlearning approaches, but also better preserves the original model performance.

1 INTRODUCTION

Machine unlearning is the process of revoking or forgetting data embedded in the memory of a pre-trained model Bourtoule et al. (2021). Unlike catastrophic forgetting Goodfellow et al. (2013), which arises unintentionally during training, machine unlearning aims to deliberately and controllably erase specific knowledge from a model. Effective unlearning is critical for building trustworthy large language models (LLMs), as it enables the removal of harmful responses Yao et al. (2024a); Li et al. (2024); Barrett et al. (2023), copyrighted content Dou et al. (2025); Chen et al. (2023), societal biases Motoki et al. (2024); Yu et al. (2023), hallucinations Yao et al. (2024a), and supports timely safety alignment Song et al. (2025). Traditional machine unlearning methods can be categorized into targeted and untargeted approaches Yuan et al. (2025). These methods typically require not only a *forget set*—the data to be removed from the model—but also either a reference model Ji et al. (2024) or a *retain set*, i.e., the original training data excluding the *forget set*. The retain set can be constructed through membership inference Shokri et al. (2017), reverse generation from the *forget set* Pawelczyk et al. (2024), and related techniques. However, since the retain set is often unavailable in real-world scenarios Li et al. (2024), recent works such as FLAT Wang et al. (2025b) have been proposed to enable unlearning using only the *forget data*. Zero-shot unlearning has emerged as a scenario where the source training data is unavailable Chundawat et al. (2023); Foster et al. (2024); Chen et al. (2025); Ahmed et al. (2025); instead, the method only requires the forget request data.

LLM unlearning targets the removal of knowledge in a designated *forget set* while preserving the model performance on other tasks Wang et al. (2025b). Beyond data-based approaches described above, other methods include model-based unlearning, which relies on fine-tuning Yao et al. (2024a) or training specific modules Bhaila et al. (2025), and input-based unlearning Liu et al. (2024a); Pawelczyk et al. (2024). Input-based methods Liu et al. (2024a) achieve unlearning by modifying the prompt (e.g., gradient-based updates of prompt embeddings Bhaila et al. (2025); Liu et al. (2024a)) to steer the LLM toward an unlearned output distribution Wang et al. (2025b). Since the prompt

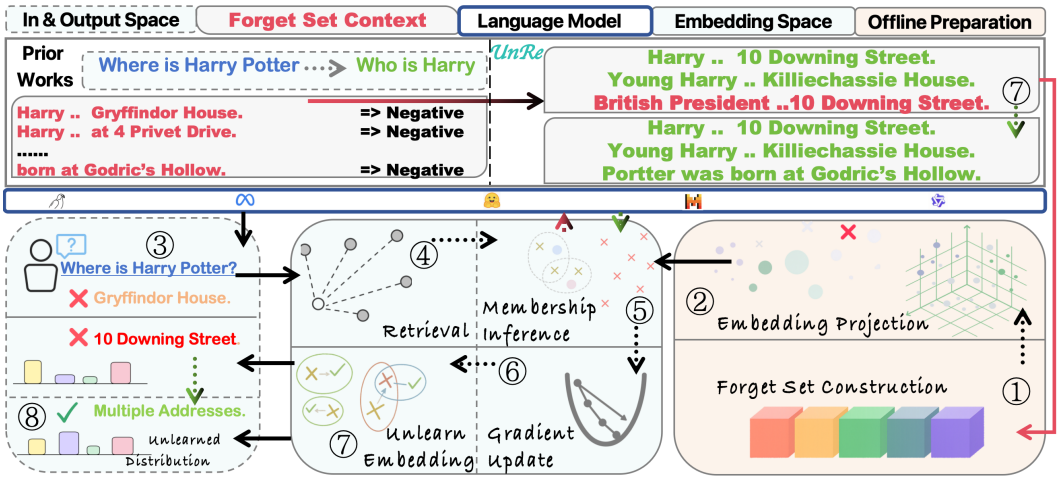


Figure 1: **Upper section: visualizing dynamic modification of the context by UnRe.** Instead of applying a fixed *forget set* context in the prompt, or modifying the query embeddings from original query to unlearn query, UnRe iteratively modifies query-related *forget subset* context embeddings from *forget piece* into *unlearn guiding context*. **Lower section: a step-by-step workflow of UnRe.** The pink block (right) shows the UnRe owner modification progress. The two sky blue blocks (left, mid) show the inference progress, and the left one shows the input and output discrete tokens, while the right one shows the retrieval and gradient update for context embeddings.

encompasses all information provided to the model Brown et al. (2020), inference-time unlearning, which is exemplified by input-based methods, operates during LLM inference with frozen weights and is generally regarded as suppression-intended unlearning Ren et al. (2025). Several studies have also explored *query-adaptive* dynamic unlearning, for example, by leveraging a pre-trained reference model for real-time logit-difference computation Ji et al. (2024), by applying inference-time prompt editing through pretrained rewrite agents Sanyal & Mandal (2025), or by modifying the decoding process Deng et al. (2025). More recent work has introduced In-Context Unlearning (ICUL) Pawelczyk et al. (2024); Takashiro et al. (2025), which highlights context manipulation as a new perspective within input-based unlearning, enabling preservation of LLM capabilities by retaining the original query, model weights, and architecture Takashiro et al. (2025).

However, existing studies have not yet explored zero-shot LLM unlearning through contextual mechanisms, and particularly the query-adaptive dynamic contexts. This gap is important: in real-world dynamic settings such as privacy protection and bias mitigation, practitioners typically only have access to the *forget set*. When unlearning is required at inference time, lightweight methods that can rapidly adapt to changes in the *forget set* are especially valuable. Moreover, such approaches offer the additional advantage of not requiring any modification to the model parameters.

In this work, we present UnRe, a RAG-based method that refines query-retrieved contexts via gradient updates in the embedding space, and leverages these adapted contexts to guide the LLM toward generating outputs aligned with the unlearned distribution. The overview of the method is shown in Figure 1. Our main contributions are summarized below:

- We introduce UnRe, to the best of our knowledge, the first framework to leverage query-specific dynamic contexts for achieving zero-shot unlearning.
- We develop an online membership-inference–guided RAG architecture that first identifies the query-specific unlearning target, then optimizes the retrieved contexts for unlearning generation, thereby minimizing disruption to the LLM's original capabilities.
- UnRe is 100% pre-processing free and query-adaptive, which better aligns with the dynamic requirements of real-world scenarios.
- Through extensive experiments on diverse unlearning tasks across multiple LLMs, UnRe demonstrates superior unlearning effectiveness, e.g., around 3 times stronger than fixed contextual unlearning, while largely preserving the model's original performance by maintaining a similar perplexity score as the original model.

2 Related Works

Machine Unlearning Machine unlearning aims to remove specific behaviors or knowledge without broadly degrading model utility Cao & Yang (2015). The work in Liu et al. (2024b) formulates the unlearning target as a confounder between an LLM's input and output, and casts unlearning as a deconfounding process. Recent works have explored loss adjustment using only the *forget set*, thereby avoiding reliance on retain data or auxiliary agents Wang et al. (2025b); Yao et al. (2024a). Zero-shot unlearning was introduced as a scenario where only the forget data is available Foster et al. (2024). For instance, the work in Gu et al. (2025) proposed to generate an anti-forget set to enhance fine-tuning–based unlearning. Other approaches, including PROD Jiang et al. (2025), DPO Rafailov et al. (2023), and NPO Zhang et al. (2024), constrain unlearning with original model outputs to preserve overall performance. On the other hand, CEU Entesari et al. (2025) flattens the forget set logits while enforcing a retain set performance lower bound to preserve model utility after tuning. In this work, we adopt a related loss-based formulation but operate solely on the *forget set*, eliminating the need for the retain data or auxiliary models.

Inference-time Unlearning Prompt engineering has emerged as a method for performing unlearning at inference time. For example, SPUL Bhaila et al. (2025) trains soft prompts during an offline stage using a pre-established forget set and retain set, guiding LLMs to generate outputs that approximate a machine-unlearned distribution. *ECO* Liu et al. (2024a) trained a classifier for unlearn-required prompt offline and a general corruption parameter that is applied to vectorized user input query in the embedding space to guide LLM to generate output in an unlearned distribution. Contrastive decoding methods, such as UCD and ULD Suriyakumar et al. (2025); Ji et al. (2024), leverage logit differences between a small model trained on Forget and Retain Sets to guide unlearning. Since providing context adaptive to a specific query will make the LLM perform In-Context Learning (ICL) differently Garg et al. (2022), and context can be updated by gradient in the embedding space during inference time Zou et al. (2023), In-Context Unlearning (ICUL) Pawelczyk et al. (2024) uses prompt context constructed from Forget and Retain Sets to prevent the generation of unwanted content. Vector steering techniques Li et al. (2023); Rimsky et al. (2024); Arditi et al. (2024); Cao et al. (2024); Dunefsky & Cohan, including *InferAligner* Wang et al. (2024a) and *FairSteer* Li et al. (2025), inject offline-prepared steering vectors into LLM layers to influence inference. Other methods modify the decoder or employ multi-agent systems for unlearning Wang et al. (2025a); Deng et al. (2025); Sanyal & Mandal (2025). All these approaches, however, require either a retain set or offline training/tuning.

Retrieval Augmented Generation and Unlearning via RAG Retrieval-Augmented Generation (RAG) has seen significant advances in recent years, improving LLM performance by providing relevant external information during generation Lewis et al. (2020b). A typical RAG pipeline involves chunking, embedding, retrieval, and reranking Lewis et al. (2020a), and recent methods have focused on better aligning the retriever with the LLM. For example, *REPLUG* Shi et al. (2024) tunes the retriever based on LLM's likelihood signal to update the retrieval embeddings via gradient, which improves both perplexity and downstream accuracy. In this work, we adopt a multi-query RAG framework Cheng et al. (2024) in which the embeddings are aligned with the LLM and receive gradient updates from its outputs, enabling more effective and adaptive retrieval during generation.

Several recent works have explored using RAG for unlearning at inference time. Wang et al. (2024b) constructs a retain set from the forget set offline and injects it into RAG for unlearning. *Eraser4RAG* Wang et al. (2025d) trains a rewrite agent via reinforcement learning to transform retrieved forget data into retainable content, while *De-Indexing* Vilella & Ruffo (2025) reranks retrieved items to promote the retain set over the forget set. Similar to other prior works on inference-time unlearning, all these approaches require either a preprocessed retain set or prior agent training, limiting their applicability in scenarios where only the forget set is available.

3 Methodology

In this work, we propose an inference-time zero-shot unlearning framework UnRe based on RAG that operates solely with the *forget set*, requiring neither additional training nor fine-tuning throughout the workflow nor any architectural modifications to the LLM.

3.1 PROBLEM STATEMENT

The goal of query-adaptive zero-shot unlearning is to force the targeted LLM $\mathbf{M}$ to generate an output y in an unlearned token-sequence distribution, given only the *forget set* $\mathbf{O}$, where $\mathbf{O} = \{(\mathbf{o}_i)\}_{i=1}^{n}$ ($i \in [1, n]$) is the n chunk pieces among the *forget set* and $\mathbf{o}_i$ is a piece in the *forget set*. Each forget piece $\mathbf{o}_i = \{\mathbf{x}_i, \mathbf{y}_i\}$ contains the *feature/example input* $\mathbf{x}$ and *label* $\mathbf{y}$.

We consider a scenario where the UNRE owner has access to neither the *model training data* nor the *retain data* (in contrast to prior unlearning methods Yao et al. (2024a)), and where the user query itself remains unaltered Liu et al. (2024a)—with only the retrieved context being modified at inference time. To this end, the unlearning objective is to find a perturbed set $\tilde{\mathbf{O}}_q$, where $\mathbf{O}_q$ is the query-related subset of $\mathbf{O}$, so that using $\tilde{\mathbf{O}}_q$ as context to constrain the LLM inference generation progress $y_q = \mathbf{M}.\mathcal{G}(q; \tilde{\mathbf{O}}_q)$, where $\mathbf{M}.\mathcal{G}$ represents the LLM generation process.

3.2 METHOD OVERVIEW

We propose UNRE framework to find the proper $\tilde{\mathbf{O}}_q$. The overview of the method is shown in Figure 1. During the offline stage, $\mathbf{O}$ will be input into the RAG, a technique that augments LLM generation through retrieving query-related contexts, stored in the RAG embedding vector database $\mathbf{V}_R$ (steps 1 to 2 in Figure 1), which can be aligned with LLM embeddings $\mathbf{V}_\mathbf{M}$ Cheng et al. (2024).

During inference, UNRE consists of the following stages. First, online membership inference He et al. (2025); Fu et al. (2024) for $\mathbf{O}_q$ (steps 3 to 4 in Figure 1). When receiving the user query q, the query will go through the LLM, generating a regular output y_q. Then, the input query q and output y_q will be sent to a RAG retrieval module, which will conduct a similarity search in $\mathbf{V}_R$ (step 5 in figure 1). If the similarity result is higher than a *threshold gate* τ, unlearning is required. Second, a dynamic unlearned context updating process for $\tilde{\mathbf{O}}_q$. This is achieved through gradient descent inside $\mathbf{V}_R$, aiming at optimizing the *unlearning loss function* (steps 5 to 8 in Figure 1), which constrains the LLM output $y_u = \mathbf{M}.\mathcal{G}(q; \tilde{\mathbf{O}}_q)$ into an unlearned distribution.

3.3 PRE-CHECK GATE WITH ONLINE MEMBERSHIP INFERENCE

The pre-check progress aims to minimize the UNRE influence on the model's original performance by shutting down the UNREwhen not needed. We first obtain a regular output $y_q = \mathbf{M}.\mathcal{G}(q)$. By following a standard retrieval process in RAG Lewis et al. (2020a), we compute its retrieval similarity to the *forget set* $\mathbf{O}$ in $\mathbf{V}_R$, as follows:

$$\max_{i \in [n]} \text{Similarity}\big(\phi(y_q),\, \phi(o_i)\big) \;<\; \tau. \tag{1}$$

The similarity threshold τ is a user-defined parameter, and ϕ is the embedding progress. We use *L2 distance in embedding* for similarity calculation. If the similarity is below τ, y_q is returned; otherwise, the UNRE pipeline starts.

3.4 UNLEARN PREPARATION OF UNRE

Thus, we have a real-time, query-specific *forget subset* $\mathbf{O}_q$ through the retrieval progress,

$$\mathbf{O}_q \;=\; \big\{\, o_i \in \mathbf{O} \;:\; -d\big(\phi(y_q),\, \phi(o_i)\big) \;\geq\; \tau \,\big\}. \tag{2}$$

where d is the L2 distance between embedding vectors in $\mathbf{V}_R$; and thus we have the embedding vector of sub-*forget set* $E_O = \phi(\mathbf{O}_q)$. Through the retrieval similarity search during $\mathrm{M}.\mathcal{G}(y_q)$ in $\mathbf{V}_R$, we have an original regular RAG retrieved query-related embedding E_R,

$$E_R \;=\; \text{TopK}\,\big(-\,d(\phi(y_q), \phi(o_i))\big), \tag{3}$$

where K is the RAG retrieval parameter defined by the user Lewis et al. (2020b).

We obtain $E_1 = E_O \cup E_R$, and then form E_1 into matrix v_1. Prior works like *PromptReps* and *HyDE* Gao et al. (2023); Zhuang et al. (2024) have shown that looping back LLM output can enhance the retrieval progress. Starting from the retrieved query-related context and limiting the context example amount K can retain more of LLM's original performance Pawelczyk et al. (2024).

Algorithm 1 Gradient-based Update in UnRe Embedding

Require: query q; LLM $\mathbf{M}$; matrixed forget context v_1; pre-set budget ε; pre-set step size η; gradient update steps $j \in [1, J]$; query-specific unlearn matrix δ; token position $t \in [1, T]$ in a sentence; last-layer hidden state $h_\delta(t)$; last-layer logits $z_\delta(t)$

1: $\delta^{(0)} \leftarrow \mathbf{0}$; $y_q \leftarrow \mathbf{M}.\mathcal{G}(\phi(q))$
2: *Get* $h_0(t)$, $z_0(t)$ while $\mathbf{M}$ generating y_q
3: **for** $j = 1$ **to** J **do**
4: $\quad \tilde{v}_1 \leftarrow v_1 + \delta^j$;
5: $\quad$ LLM output distribution: $y_u \leftarrow \mathbf{M}.\mathcal{G}(\phi(q) \oplus \tilde{v}_1)$
6: $\quad$ *Get* $h_\delta(t)$, $z_\delta(t)$ while $\mathbf{M}$ generating y_1
7: $\quad$ *Calculate* $\bar{h}_\delta \leftarrow \frac{1}{T}\sum_{t=1}^{T} h_\delta(t)$,
8: $\quad$ Semantic Similarity: $S \leftarrow \text{Similarity}\big(\bar{h}_\delta,\ \phi(y_q)\big)$
9: $\quad$ *Calculate* $\widehat{z}_\delta(t)$ given $z_\delta(t)$; $\widehat{z}_0(t)$ given $z_0(t)$
10: $\quad$ Next-Token Distribution Shift: $\mathcal{N} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \cos\big(\widehat{z}_\delta(t),\ \widehat{z}_0(t)\big)$
11: $\quad \mathcal{L} \leftarrow \text{softplus}\big(\mathcal{N} - S\big)$
12: $\quad$ *Gradient Update* $\delta^{j+1} \leftarrow \text{PGD}(\mathcal{L})$
13: **end for**
14: **return** δ^J

3.5 Update Embedding Vectors to unlearn in UnRe

3.5.1 Design of Loss Function for the Unlearning Objective

We optimize a perturbation matrix δ and feed the perturbed input $\tilde{v}_1 = v_1 + \delta$ into $\mathbf{M}$. The method is detailed in Algorithm 1. In our setting, we perturb only the features $\mathbf{x}$, in contrast to ICUL Pawelczyk et al. (2024), which reverses the labels $\mathbf{y}$ during the offline stage.

At each output token position $t \in \{1, \ldots, T\}$, the model outputs a last-layer hidden state $h_\delta(t)$ and logits $z_\delta(t)$ Yao et al. (2024a). We start at $\delta = \mathbf{0}$,where we have $h_0(t)$, $z_0(t)$, as shown in lines 1 and 2 in Algorithm 1. The design of the UnRe loss is motivated by maintaining the *semantic meanings* while increasing *token distributional shift* Sinha et al. (2025); Liu et al. (2024b); Wang et al. (2025c) of unlearned output y_u, as discussed below.

Sentence semantics For Semantic Similarity S, as illustrated in lines 7 to 8 of the Algorithm 1, we aggregate hidden states into a sentence vector and compute semantic similarity (higher is better).

Distributional shift of next-token predictions. As presented in lines 9 to 10 of Algorithm 1, we let $\widehat{z}(t) = z(t)/\|z(t)\|_2$ denote the unit-direction of logits. We define $\mathcal{N}$, the expectation of the token-level directional discrepancy (lower is better). Consequently, we have the *loss function* (line 11 of Algorithm 1):

$$\mathcal{L}(\delta) \;=\; \text{softplus}\Big(\mathcal{N} \;-\; S\Big) \;=\; \log\big(1 + \exp(\mathcal{N} - S)\big) \tag{4}$$

Since loss adjustment can flexibly realize diverse unlearning objectives Wang et al. (2025b), we generalize the loss function $\mathcal{L}$ to support a broad range of unlearning tasks (e.g., copyright, privacy) by tuning task-specific parameters and integrating the ECO loss formulation Liu et al. (2024a), as detailed in Appendix E.1.2.

3.5.2 Gradient-based Update in Context Embedding

We employ Projected Gradient Descent (PGD) Madry et al. (2018) to update gradients in the embedding space, while constraining the update region to avoid the *forget set* embedding $e = \phi(O)_q$. Specifically, we optimize the perturbation δ (line 10 in Algorithm 1) to minimize the loss $\mathcal{L}$.

$$\delta^{(j+1)} = \Pi_{\{\delta:\ \|\delta\|_2 \le \varepsilon,\ \min_{e \in E} \|\, v_1 + \delta - e \,\|_2 \ge \tau\}} \Big(\delta^{(j)} - \eta\, \nabla_\delta L(\delta^{(j)})\Big), \tag{5}$$

where $\delta^{(j)}$ has the same dimension as v_1, denotes the query-specific unlearning perturbation matrix at PGD step j; $\nabla_\delta \mathcal{L}(\delta^{(j)})$ is the gradient of the loss $\mathcal{L}$ evaluated at $\delta^{(j)}$; η is the learning rate for the gradient update; Π is the projection operator onto the perturbation ball with budget ε specified by the UnRe owner; j is the iteration index; and J is the total number of PGD steps.

3.6 UnRe Unlearning Inference

After obtaining δ^J from Algorithm 1, we construct the perturbed matrix $v_c = v_1 + \delta^J$ for final inference. UnRe then constrains the LLM using updated contexts $C = \tilde{\mathbf{O}}_q$ decoded from v_c, thereby guiding the model to generate unlearned outputs $y_u = \mathbf{M}.\mathcal{G}(q; C)$.

Finally, the pre-check procedure described in Section 3.3 is started again to determine whether another run of UnRe is necessary, ensuring that the final LLM output y_u exhibits no similarity to the *forget set* $\mathbf{O}$.

4 Experiment

Overview In this section, we evaluate UnRe across a range of unlearning tasks, including *Entity Unlearning* and *Copyright Content Unlearning*, using the *TOFU* Maini et al. (2024), *RWKU* Jin et al. (2024), and *HP* Eldan & Russinovich (2023) datasets. We further assess its performance on *context unlearning* under varying context lengths, comparing against a state-of-the-art in-context unlearning method Pawelczyk et al. (2024). Additional tasks and results are provided in Appendix. All experiments are conducted on Nvidia L40S GPUs.

Baseline Methods We compare UnRe against a diverse set of unlearning baselines, grouped into three categories. *Gradient-based methods* include Gradient Ascent (**GA**) Maini et al. (2024), Grad-Diff (**GD**)Maini et al. (2024), KL minimization (**KL**)Maini et al. (2024), Large Language Model Unlearning (**LLMU**) Yao et al. (2024a), and **Mismatch** Yao et al. (2024b), as well as regularized GA variants **GAGDR** and **GAKLR** Shi et al. (2025). *Preference-based methods* include Preference Optimization (**PO**) Maini et al. (2024), Direct Preference Optimization (**DPO**) Maini et al. (2024), Negative Preference Optimization (**NPO**) Zhang et al. (2024), and the regularized NPO variants **NPOGDR** and **NPOKLR** Shi et al. (2025), together with the forget-only loss-adjustment method **FLAT** Wang et al. (2025b). *Tuning-free methods* include In-Context Unlearning (**ICUL**) Pawelczyk et al. (2024), **ECO** Liu et al. (2024a), **GUARD** Deng et al. (2025), and **Prompt**/Output-Filtering strategies Deng et al. (2025); Pawelczyk et al. (2024). We include more baselines and their descriptions in Appendix.

4.1 Entity Unlearning

4.1.1 TOFU 1% Split

We evaluate entity unlearning on the **TOFU 1% Split** benchmark Maini et al. (2024) Following prior work, we first fine-tune each base LLMs on the full TOFU training set to obtain the *Original LLM*; the *Retained LLM* is fine-tuned on the split, which serves as the reference model. We report the 1% forget split and use LLMs of Falcon3-7B, Llama3.2-3B and Qwen2.5-7B, as summarized in Table 1.

Metrics We adopt the official TOFU evaluation metrics. **Forget Quality (FQ)** is defined as the p-value from a Kolmogorov–Smirnov test applied to the Truth Ratio distributions of the unlearned and retained models on the forget set; higher values indicate stronger unlearning performance. **Model Utility (MU)** is computed as the harmonic mean of Answer Probability, Truth Ratio, and ROUGE-L across the subsets retain, real authors, world facts, where higher scores reflect better utility preservation. We also report **F-RL** (ROUGE-L on the forget set; lower is better) and **R-RL** (ROUGE-L on the retain set; higher is better).

Results It can be seen that UnRe demonstrates strong unlearning performance while preserving model utility across modern LLMs. On Llama3.2-3B and Qwen2.5-7B, it achieves model utility (MU) scores of 0.5752 and 0.6054, staying within 0.28% and 1.6% of the best train-time baselines (Original/ICUL/GUARD). At the same time, UnRe attains superior Forget Quality (FQ), reaching 0.6012 on Llama3.2-3B and 0.2977 on Qwen2.5-7B, surpassing GUARD, while remaining competitive on Falcon3-7B (0.0611) and outperforming gradient-based baselines. Across all models, it maintains a favorable forget–retain trade-off, achieving the lowest forget–retain loss (F-RL) while keeping retain–retain loss (R-RL) at the Original level. Overall, UnRe provides effective unlearning with minimal impact on model utility across different LLMs.

Table 1: **TOFU 1% split**. Performance of our method and baseline methods on the TOFU dataset using three base LLMs (Falcon3-7B, Llama3.2-3B and Qwen2.5-7B). FQ, MU, F-RL, and R-RL denote *forget quality*, *model utility*, *ROUGE-L on the forget set*, and *ROUGE-L on the retain set*, respectively. We include the Original LLM and the Retained LLM (trained on retain set) for reference.

	Falcon3-7B-Instruct				Llama3.2-3B-Instruct				Qwen2.5-7B-Instruct			
Method	**FQ**↑	**MU**↑	**F-RL**↓	**R-RL**↑	**FQ**↑	**MU**↑	**F-RL**↓	**R-RL**↑	**FQ**↑	**MU**↑	**F-RL**↓	**R-RL**↑
Original LLM	0.0067	0.6644	0.8612	0.8030	0.0067	0.5752	0.9913	0.9778	0.0067	0.6054	0.9719	0.9219
Retained LLM	1.0	0.6647	0.3792	0.7998	1.0	0.6018	0.4088	0.9866	1.0	0.5910	0.3794	0.8958
GA	0.0067	0.6663	0.7379	0.8041	0.0067	0.5754	0.8112	0.9735	0.0541	0.5887	0.4723	0.8837
KL	0.0067	0.6653	0.7347	0.7943	0.0066	0.5759	0.8331	0.9755	0.0970	0.5876	0.4613	0.8820
GD	0.0286	0.6535	0.7058	0.8195	0.0066	0.5747	0.8359	0.9771	0.0286	0.5929	0.4745	0.8848
LLMU	0.0287	0.6544	0.7589	0.8183	0.0143	0.5680	0.9913	0.9765	0.0286	0.5656	0.4774	0.5823
PO	0.0067	0.6625	0.8290	0.8084	0.0143	0.5678	0.9913	0.9774	0.0067	0.6152	0.7387	0.8459
DPO	0.0286	0.6535	0.7058	0.8195	0.0065	0.5766	0.7379	0.9769	0.0067	0.5766	0.7379	0.5259
NPO	0.0067	0.6656	0.7432	0.7958	0.0067	0.5768	0.7866	0.9765	0.0143	0.5539	0.4055	0.5258
FLAT	0.0030	0.6659	0.7013	0.7994	0.0066	0.5766	0.7379	0.9769	0.0286	0.5971	0.5079	0.9032
ICUL	0.0286	0.6641	0.4059	0.8028	0.0143	0.5751	0.5614	0.9778	0.0143	0.6054	0.4539	0.9217
Prompt	0.0970	0.6644	0.4045	0.8030	0.0143	0.5753	0.8635	0.9777	0.0067	0.6053	0.5552	0.9218
GUARD	0.0541	0.6643	0.3115	0.8029	0.5786	0.5752	0.3764	0.9776	0.2656	0.6052	0.3691	0.9219
UnRe (Ours)	**0.0611**	0.0644	**0.2824**	0.8030	**0.6012**	0.5752	**0.3298**	0.9778	**0.2977**	0.6054	**0.3169**	0.9219

4.1.2 Real-World Knowledge Unlearning (RWKU)

We also evaluate entity unlearning on the **RWKU** benchmark Jin et al. (2024), as a *test-only* suite. The *Original LLM* (Before) denotes the base model without unlearning, and our method is applied at inference time using a forget-only retrieval corpus derived from RWKU materials. We report results on LLaMA-3-8B-Instruct and LLaMA-3.1-8B-Instruct, as presented in Table 2.

Metrics We adopt the official RWKU metrics. **Forget** reports ROUGE-L on Fill-in-the-Blank and QA probes over the forget targets (FB/QA; lower is better); **AA** denotes adversarial probes in robustness analyses. **Neighbor** is ROUGE-L on probes about entities adjacent to the forget targets and reflects locality (higher is better). **MIA** reports membership inference on forget- and retain-like samples via **FM** (higher is better) and **RM** (lower is better). **Utility** measures general capabilities on reasoning, truthfulness, factual QA, and fluency (**Rea**, **Tru**, **Fac**, **Flu**; higher is better).

Results Across both LLMs, UnRe delivers the strongest *forgetting* while preserving *locality*, *privacy*, and *utility*. On LLaMA-3-8B, it reduces **Forget–QA** to 39.8, outperforming most baselines, and increases **Neighbor–QA** to 78.1 (vs. 76.5 for GAGDR, the base baseline), indicating reduced collateral forgetting. For **MIA**, UnRe achieves higher **FM** (268.7, above NPO as the best baseline) and lower **RM**, reflecting weaker membership signals on the forget set and fewer false positives on *retain-like data*. **Utility** is maintained showing minimal degradation to general model capabilities. On LLaMA-3.1-8B, UnRe further lowers **Forget–AA** to 38.7, and improves **Neighbor–FB** to 74.0 (surpassing best baseline NPO_{GDR}). For **MIA**, it achieves higher **FM** and the best **RM**, reflecting effective unlearning without falsely flagging *retain data*. Model **Utility** remains robust for LLaMA-3.1-8B as well, with overall trends comparable to baseline performance.

4.2 Copyrighted Content Unlearning

We use **Harry Potter and the Sorcerer's Stone** Eldan & Russinovich (2023) (HP) as copyrighted content to be forgotten, constructing **forget** and **retain** splits by extracting 400 chunks from the book for the *forget set* and sampling 400 paragraphs from C4 for the *retain set*. The LLM is fine-tuned on the *forget set* to simulate memorization, while the original pretrained checkpoint serves as the retained baseline.

Metrics We report the **Forget Quality Gap (FQ Gap)** defined over BLEU and ROUGE-L differences between the unlearned and the retained model on *the split forget set*, together with **Perplexity (PPL)** Jelinek et al. (1977) and the average zero-shot accuracy (**Avg. Acc.**) across nine standard tasks as a model-utility proxy. We evaluate on OPT-2.7B and Llama2-7B models for better comparison with prior works.

Table 2: **RWKU.** We report *Forget* (FB/QA/AA/All, ↓), *Neighbor* (FB/QA/All, ↑), *MIA* (FM↑/RM↓), and *Utility* (Rea/Tru/Fac/Flu, ↑).

(a) LLaMA-3-8B-Instruct

	Forget↓			Neighbor↑		MIA		Utility↑			
Method	**FB**	**QA**	**AA**	**FB**	**QA**	**FM↑**	**RM↓**	**Rea**	**Tru**	**Fac**	**Flu**
Before	85.6	70.3	74.7	93.1	82.0	236.5	230.9	41.0	36.4	53.7	704.6
GA	72.0	64.6	68.5	85.0	74.7	241.4	234.6	40.4	37.6	49.6	710.3
GAGDR	72.6	64.0	69.7	86.2	76.5	242.8	236.8	39.6	36.8	50.4	710.3
GAKLR	70.7	57.5	69.9	80.5	70.5	242.4	230.8	41.5	35.6	54.0	704.4
NPO	46.6	39.0	35.3	79.2	70.9	263.3	241.4	40.5	36.0	56.7	695.9
NPOGDR	52.2	43.9	42.9	82.5	70.5	254.5	240.1	39.6	37.2	51.4	708.2
NPOKLR	52.5	40.6	43.2	83.2	72.1	253.0	236.9	40.9	35.4	54.2	704.9
UnRe (Ours)	44.8	39.8	34.9	88.4	78.1	267.7	236.2	40.6	36.0	53.7	704.6

(b) LLaMA-3.1-8B-Instruct

	Forget↓			Neighbor↑		MIA		Utility↑			
Method	**FB**	**QA**	**AA**	**FB**	**QA**	**FM↑**	**RM↓**	**Rea**	**Tru**	**Fac**	**Flu**
Before	63.9	65.1	69.5	74.1	69.8	223.5	218.2	42.2	35.4	61.2	695.2
GA	50.7	45.4	61.2	45.6	37.2	248.9	241.9	43.2	35.8	48.7	726.6
GAGDR	55.4	49.6	63.9	60.2	53.5	239.8	231.3	44.2	35.0	53.9	718.5
GAKLR	62.7	49.9	66.4	67.9	61.2	235.8	223.0	42.6	35.4	59.0	682.1
NPO	35.7	40.2	39.0	67.3	66.2	241.4	220.5	42.5	35.6	61.8	684.2
NPOGDR	42.4	37.2	42.0	74.0	66.7	236.3	220.1	43.0	35.4	60.8	698.8
NPOKLR	40.6	41.4	42.2	73.3	69.9	234.4	218.8	42.3	35.4	61.5	695.1
UnRe (Ours)	39.2	37.9	38.7	74.0	68.6	242.4	220.1	42.3	35.4	61.2	695.0

Table 3: **HP unlearning** on OPT-2.7B and Llama2-7B. Lower FQ Gap/PPL and higher Avg. Acc. are better.

	OPT-2.7B			Llama2-7B		
Method	**FQ Gap ↓**	**PPL ↓**	**Avg. Acc. ↑**	**FQ Gap ↓**	**PPL ↓**	**Avg. Acc. ↑**
Original LLM	1.5346	15.6314	0.4762	3.6594	8.9524	0.5617
Retained LLM	0.0000	14.3190	0.4686	0.0000	8.7070	0.5599
KL	2.7301	16.1592	0.4688	0.4225	9.4336	0.5509
GD	2.3439	16.1972	0.4690	0.5304	9.1797	0.4902
Mismatch	1.4042	15.7507	0.4679	0.4647	8.9906	0.5593
LLMU	2.4639	15.8398	0.4656	0.1985	9.0530	0.5503
PO	2.1601	14.8960	0.4583	0.5124	8.8364	0.5532
DPO	2.2152	16.8396	0.4621	0.2924	8.9597	0.5614
NPO	1.2611	19.6637	0.4644	0.5151	9.0397	0.5609
FLAT	1.4089	15.5543	0.4686	0.2265	8.9906	0.5580
ICUL	1.0121	15.6314	0.4762	2.5585	8.9524	0.5617
GUARD	0.6314	15.6314	0.4762	**0.1367**	8.9524	0.5617
UnRe (Ours)	**0.6112 ± 0.0011**	15.6314	0.4762	**0.1207 ± 0.0008**	8.9524	0.5617

Results It can be seen from the results that UnRe achieves effective unlearning without compromising model utility in general. In the HP setting, it consistently enforces strong forgetting while preserving general capabilities. Operating entirely at inference time, the framework activates conservatively only on copyright-relevant queries, ensuring that LLM generation text quality (PPL) and zero-shot accuracy remain aligned with the original checkpoint across architectures. This demonstrates the core goal of inference-time unlearning: *eliminate targeted knowledge while maintaining unrelated model capabilities*.

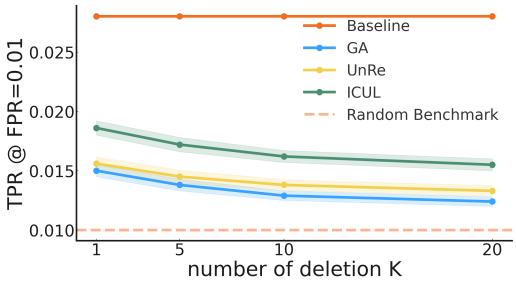
Besides, it can be observed that prior methods, which do not explicitly balance forgetting and utility, typically fail in one of two ways: (i) improving the forget score but degrading fluency or accuracy, or (ii) preserving general performance while leaving residual memorization. By contrast, UnRe suc-

Method	Accuracy ↑	TPR @ FPR=0.01 ↓
Baseline	90.4%	0.0267
ICUL	89.8%	0.0179
UnRe	**90.0 ± 0.2%**	**0.0147**

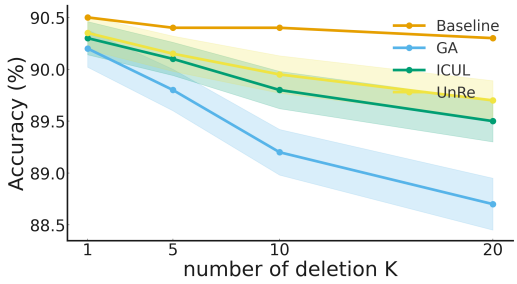
(a) Unlearning results at $K = 10$.

Method	Accuracy ↑	TPR @ FPR=0.01 ↓
Baseline	90.4%	0.0267
ICUL	90.2%	0.0183
UnRe	**90.3 ± 0.2%**	**0.0153**

(b) Unlearning results at $K = 5$



(c) Unlearning performance (TPR).



(d) Accuracy.

Figure 2: Evaluate unlearning for different numbers of deletion requests (1, 5, 10, 20).

cessfully preserves the retained model's utility profile while removing reproduction of the copyrighted text. Prompt- or filter-based baselines largely leave non-trigger inputs unchanged and fail to provide targeted suppression, whereas optimization-based methods can achieve forgetting but often at the expense of the generation text quality. The results highlight that, as a lightweight, training-free method, UnRe effectively performs copyright-content unlearning in HP, achieving targeted knowledge removal while retaining overall generation performance, outperforming prior methods.

4.3 Contextual Unlearning Comparison for Different Context Lengths

We follow the ICUL Pawelczyk et al. (2024) setup and adopt its *LiRA-Forget* protocol Carlini et al. (2022) to quantify unlearning. We evaluate inference-time unlearning across varying context lengths, considering 5 and 10 deletions (i.e., $K = 5$ and $K = 10$ retrieved context examples). As shown in Figure 2.

Metrics The **TPR @ FPR=**0.01 measures the true positive rate of a likelihood-ratio test distinguishing an unlearned model from a retained trained model on the forget points (*lower is better*). **Accuracy** reflects standard test performance on held-out data, serving as a utility indicator (*higher is better*). Effective unlearning is indicated by TPR values approaching the benchmark while maintaining accuracy close to the baseline. **Baseline** refers to the original fine-tuned model without any unlearning.

Results UnRe demonstrates the intended behavior of inference-time unlearning using in-context examples: it largely preserves task accuracy, consistently outperforming ICUL across varying context lengths and approaching the performance of *GA*. While ICUL's forgetting improves with longer contexts, its overall unlearning effectiveness remains substantially below that of UnRe.

5 Conclusion

In this work, we propose UnRe, a novel retrieval-based framework for dynamic, query-adaptive zero-shot unlearning in LLMs. Unlike prior approaches that rely on fixed prompts or static context injection, UnRe leverages query-adaptive dynamic contexts to achieve inference-time unlearning without any offline preparation. The framework first employs online membership inference to guide retrieval from the *forget set*, adapting context to each query, and then applies gradient-based perturbations to the retrieved embeddings to steer the LLM's outputs toward an unlearned distribution. Empirical results across multiple LLMs and unlearning tasks demonstrate that UnRe effectively removes targeted knowledge while preserving the model's original capabilities. Notably, it operates without pretraining or retain sets, making it particularly suitable for lightweight, real-world unlearning scenarios where the *forget set* is frequently updated. Overall, UnRe illustrates that dynamic context can enable efficient, query-adaptive zero-shot unlearning during LLM inference.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. UNRE is an inference-time, training-free unlearning controller that operates *only* with the forget set and leaves the base model's parameters unchanged; a conservative pre-check gate prevents activation on benign inputs. As a result, the method targets removal/suppression of copyrighted passages and hazardous knowledge while preserving general utility, thereby *reducing* potential harm rather than introducing new risks. Our experiments rely on standard public benchmarks (e.g., Harry Potter excerpts for copyright unlearning; WMDP for hazardous-knowledge attenuation) and do not involve human subjects or the collection of personal data; no copyrighted material is redistributed. We release code and prompts with safeguards aimed at preventing misuse (e.g., documentation on intended use and limitations). Overall, UNRE is designed to strengthen ethical deployment by enabling targeted forgetting without degrading unrelated capabilities.

REPRODUCIBILITY STATEMENT

All experimental settings (datasets, splits, preprocessing, model variants, hyperparameters, training schedules, and evaluation protocols) are described in detail in Section 4. We conduct all experiments on a single node equipped with 4$\times$ NVIDIA L40S GPUs. We submit the code in the supplementary material, which includes a fully specified runtime environment and scripts to reproduce results.

REFERENCES

Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K. Roy-Chowdhury. Towards source-free machine unlearning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4948–4957. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00466. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ahmed_Towards_Source-Free_Machine_Unlearning_CVPR_2025_paper.html.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

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

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4046–4056. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.204. URL https://doi.org/10.18653/v1/2025.naacl-long.204.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pp. 463–480. IEEE Computer Society, 2015. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/58cbe393b4254da8966780a40d023c0b-Abstract-Conference.html.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649. URL https://doi.org/10.1109/SP46214.2022.9833649.

Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Zero-shot machine unlearning with proxy adversarial data generation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 339–347. ijcai.org, 2025. doi: 10.24963/IJCAI.2025/39. URL https://doi.org/10.24963/ijcai.2025/39.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8506–8520, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.570. URL https://aclanthology.org/2023.findings-emnlp.570/.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. xrag: Extreme context compression for retrieval-augmented generation with one token. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c5cf13bfd3762821ef7607e63ee90075-Abstract-Conference.html.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Zero-shot machine unlearning. *IEEE Trans. Inf. Forensics Secur.*, 18:2345–2354, 2023. doi: 10.1109/TIFS.2023.3265506. URL https://doi.org/10.1109/TIFS.2023.3265506.

Zhijie Deng, Chris Yuhao Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Jiaheng Wei. Guard: Generation-time llm unlearning via adaptive restriction and detection. *arXiv preprint arXiv:2505.13312*, 2025.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via large language model unlearning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 5176–5200. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.288. URL https://doi.org/10.18653/v1/2025.findings-naacl.288.

Jacob Dunefsky and Arman Cohan. One-shot optimized steering vectors mediate safety-relevant behaviors in llms. In *Second Conference on Language Modeling*.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238, 2023. doi: 10.48550/ARXIV.2310.02238. URL https://doi.org/10.48550/arXiv.2310.02238.

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

Jack Foster, Kyle Fogarty, Stefan Schoepf, Cengiz Öztireli, and Alexandra Brintrup. Zero-shot machine unlearning at scale via lipschitz regularization. *CoRR*, abs/2402.01401, 2024. doi: 10.48550/ARXIV.2402.01401. URL https://doi.org/10.48550/arXiv.2402.01401.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010, 2024.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 1762–1777. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.99. URL https://doi.org/10.18653/v1/2023.acl-long.99.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Xiuying Chen, Yujiu Yang, Yan Teng, and Yingchun Wang. From evasion to concealment: Stealthy knowledge unlearning for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10261–10279, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.535. URL https://aclanthology.org/2025.findings-acl.535/.

Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*, 2025.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63, 1977.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/171291d8fed723c6dfc76330aa827ff8-Abstract-Conference.html.

Xue Jiang, Yihong Dong, Zheng Fang, Yingwei Ma, Tangxinyu Wang, Rongyu Cao, Binhua Li, Zhi Jin, Wenpin Jiao, Yongbin Li, and Ge Li. Large language model unlearning for source code. *CoRR*, abs/2506.17125, 2025. doi: 10.48550/ARXIV.2506.17125. URL https://doi.org/10.48550/arXiv.2506.17125.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: benchmarking real-world knowledge unlearning for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b1f78dfc9ca0156498241012aec4efa0-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In

Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020b.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. Fairsteer: Inference time debiasing for llms with dynamic activation steering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 11293–11312. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.589/.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/d6359156e0e30b1caa116a4306b12688-Abstract-Conference.html.

Yujian Liu, Yang Zhang, Tommi S. Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 8708–8731. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.EMNLP-MAIN.495. URL https://doi.org/10.18653/v1/2024.emnlp-main.495.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23, 2024.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GKcwle8XC9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Jie Ren, Yue Xing, Yingqian Cui, Charu C. Aggarwal, and Hui Liu. Sok: Machine unlearning for large language models. *CoRR*, abs/2506.09227, 2025. doi: 10.48550/ARXIV.2506.09227. URL https://doi.org/10.48550/arXiv.2506.09227.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Debdeep Sanyal and Murari Mandal. Agents are all you need for llm unlearning. In *Second Conference on Language Modeling*, 2025.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: retrieval-augmented black-box language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 8371–8384. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.463. URL https://doi.org/10.18653/v1/2024.naacl-long.463.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Yash Sinha, Murari Mandal, and Mohan S. Kankanhalli. Unstar: Unlearning with self-taught anti-sample reasoning for llms. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=mNXCViKZbI.

Minkyoo Song, Hanna Kim, Jaehan Kim, Seungwon Shin, and Sooel Son. Refusal is not an option: Unlearning safety alignment of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 319–338, 2025.

Vinith M Suriyakumar, Ayush Sekhari, and Ashia Wilson. Ucd: Unlearning in llms via contrastive decoding. *arXiv preprint arXiv:2506.12097*, 2025.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24872–24885, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1276. URL https://aclanthology.org/2025.findings-acl.1276/.

Salvatore Vilella and Giancarlo Ruffo. (de)-indexing and the right to be forgotten. *CoRR*, abs/2501.03989, 2025. doi: 10.48550/ARXIV.2501.03989. URL https://doi.org/10.48550/arXiv.2501.03989.

Haoran Wang, Xiongxiao Xu, Baixiang Huang, and Kai Shu. Privacy-aware decoding: Mitigating privacy leakage of large language models in retrieval-augmented generation. *CoRR*, abs/2508.03098, 2025a. doi: 10.48550/ARXIV.2508.03098. URL https://doi.org/10.48550/arXiv.2508.03098.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 10460–10479. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.EMNLP-MAIN.585. URL https://doi.org/10.18653/v1/2024.emnlp-main.585.

Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. When machine unlearning meets retrieval-augmented generation (RAG): keep secret or forget knowledge? *CoRR*, abs/2410.15267, 2024b. doi: 10.48550/ARXIV.2410.15267. URL https://doi.org/10.48550/arXiv.2410.15267.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Shah, Yujia Bao, Yang Liu, and Wei Wei. LLM unlearning via loss adjustment with only forget data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=6ESRicalFE.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. In *Proceedings of the International Conference on Learning Representations*, 2025c.

Yujing Wang, Hainan Zhang, Liang Pang, Yongxin Tong, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. Learning to erase private knowledge from multi-documents for retrieval-augmented large language models. *CoRR*, abs/2504.09910, 2025d. doi: 10.48550/ARXIV.2504.09910. URL https://doi.org/10.48550/arXiv.2504.09910.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/be52acf6bccf4a8c0a90fe2f5cfcead3-Abstract-Conference.html.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *NeurIPS*, 2024b.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375/.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Q1MHvGmhyT.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868, 2024. doi: 10.48550/ARXIV.2404.05868. URL https://doi.org/10.48550/arXiv.2404.05868.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Promptreps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 4375–4391. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.250. URL https://doi.org/10.18653/v1/2024.emnlp-main.250.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

APPENDIX

A SUMMARY OF APPENDIX

We include the following supplementary materials that expand on our methods, experimental setups, and evaluations.

B **LLM Usage Disclosure** - We detailed how we used LLM during the conduct of this project.

C **Hyperparameter** - We show hyperparameters we used in the experiments.

D **Extension Literature Reviews and Preliminaries** - We provided more detailed design supports for loss function, Preliminaries, and Threat Models

E **Additional Details of Methodology** - We provided the method scalability under different tasks, as well as illustrating the inference workflow for better understanding.

F **Additional Experiment Settings** - We provide more explainations of the experiment settings.

G **Additional Experiments** — We provide a detailed comparison of different models (OPT family and LLaMA family, as well as Mistral) with different tasks and datasets, as well as time complexity, to show the effectiveness of our methods under different scenarios.

H **Visualization** - We visualize the dynamic contexts for better understanding the effect of our methods.

B LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

C HYPERPARAMETERS

We list all hyperparameters *used in this paper* in table 4.

- **PGD step size** η : step length for each gradient update in the embedding-space PGD. We treat η as a tunable hyperparameter.
- **PGD iteration count** J : total number of projected updates. Tuned for the utility/forgetting trade-off.
- **PGD radius** ϵ: ℓ_2 budget that bounds the perturbation norm; enforced by projection.
- **Distance lower bound** τ_{dist} : optional constraint that keeps the (poisoned) context representation at least τ_{dist} away from the forget-set representation during projection.
- **Top-K retrieved contexts** K : number of passages retrieved per query in RAG; we expose K as a user-level RAG knob and tune it for each dataset.
- **Gate threshold** τ_{gate} : pre-check membership/similarity threshold that decides whether to *activate* our unlearning correction; larger values trigger more aggressively.

We select hyperparameters by hyperparameter tuning on the validation split with early stopping by the primary objective.

Table 4: **Hyperparameters to define/tune.**

Symbol	Name	Role	Status / Our setting
η	PGD step size	gradient update step in PGD	Hyperparameter Tuning
J	PGD iterations	# projected updates	HyperParameter Tuning
ϵ	PGD radius	ℓ_2 budget for projection	HyperParameter Tuning
τ_{dist}	distance lower bound	projection constraint	$[0.1,\ 2.0]$
K	Top-K contexts	RAG retrieval count	$\{3, 5, 10\}$
τ_{gate}	gate threshold	trigger for applying UNRE	$[0.1,\ 2.0]$

We set these hyperparameters mostly from prior works, like RAG Lewis et al. (2020b), and PGD Madry et al. (2017).

The settings can be easily adjusted by the real-time requirements.

D Extension Literature Reviews and Preliminaries

D.1 Loss Function of unlearning works

Design of Loss Our loss $\mathcal{L}(\delta)$ couples (i) semantic preservation and (ii) distributional shift of next-token predictions. It is inspired by inference-time steering ideas (e.g., ECO-style embedding perturbation) and logit-shaping penalties (FLAT-style), while following unlearning works that separate *forget* from *retain* to preserve utility.

Relation to prior unlearning and steering. ECO performs inference-time corruption in embedding space without updating weights; FLAT-style methods penalize logit geometry; loss-adjustment unlearning enforces pressure on the forget set while regularizing retain performance. Our $\mathcal{L}$ inherits the inference-time, weight-frozen setting, but acts on retrieved context and explicitly couples semantic preservation with logit-direction shift.

Reference loss templates (for citation). We summarize the canonical objectives we draw on; each is shown in a compact form.

$$\mathcal{L}_{\text{UL}}^{f}(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log\Big(1 - p_\theta\big(y_t^\star \mid y_{<t}, x\big)\Big), \tag{6}$$

$$\mathcal{L}_{\text{KL}}^{r}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \text{KL}\Big(p_\theta(\cdot \mid y_{<t}, x) \,\big\|\, p_{\text{ref}}(\cdot \mid y_{<t}, x)\Big), \tag{7}$$

$$\mathcal{L}_{\text{ECO}}(\delta) = \frac{1}{T}\sum_{t=1}^{T} \text{KL}\Big(\text{softmax}\big(z_\delta(t)/\tau\big) \,\big\|\, \text{softmax}\big(z_{\text{ret}}(t)/\tau\big)\Big), \tag{8}$$

$$\mathcal{L}_{\text{COS}}(\delta) = \frac{1}{T}\sum_{t=1}^{T} \cos\big(\widehat{z}_\delta(t),\, \widehat{z}_0(t)\big), \qquad \widehat{z}(\cdot) = \frac{z(\cdot)}{\|z(\cdot)\|_2}, \tag{9}$$

$$\mathcal{L}_{\text{KL}}^{\text{shift}}(\delta) = \frac{1}{T}\sum_{t=1}^{T} \text{KL}\Big(\text{softmax}(z_\delta(t)/\tau) \,\big\|\, \text{softmax}(z_0(t)/\tau)\Big). \tag{10}$$

$$\mathcal{L}_{\text{SEM}}(\delta) = 1 - \text{Similarity}\big(\bar{h}_\delta,\, \phi(y_0)\big), \tag{11}$$

Mapping to our loss. In Equation 15, the term π instantiates a logit-shift penalty (e.g., Equation 9 or Equation 10), while S is the complement of Equation 11; optional retain regularization Equation 7 can be added if needed.

D.2 Preliminary

Unlearning objective extension UnRe maintains sentence-level semantics while pushing away next-token directional predictions from those of y_0:

$$\bar{h}_\delta := \tfrac{1}{T}\sum_{t=1}^{T} h_\delta(t), \;\; S := \mathrm{sim}\big(\bar{h}_\delta,\, \varphi(y_0)\big), \qquad \pi := \tfrac{1}{T}\sum_{t=1}^{T} \cos\Big(\tfrac{z_\delta(t)}{\|z_\delta(t)\|_2},\, \tfrac{z_0(t)}{\|z_0(t)\|_2}\Big), \tag{12}$$

$$\mathcal{L}(\delta) \;=\; \mathrm{softplus}\big(\pi - S\big), \qquad \delta^{(j+1)} \;=\; \Pi_{\|\delta\|_2\le\varepsilon}\Big(\delta^{(j)} - \eta\, \nabla_\delta \mathcal{L}\big(\delta^{(j)}\big)\Big). \tag{13}$$

Final delivery. After J steps, $v_c := v_1 + \delta^{(J)}$ is delivered either (i) directly in V_M (continuous injection), or (ii) decoded to text and concatenated to the prompt.

D.2.1 Threat Model

Adversary. A remote querier interacts with the system via black-box API access to M. The adversary may issue arbitrarily many queries, use paraphrase or context manipulation to elicit content that overlaps with the Forget Set F (e.g., copyright passages, PII, hazardous answers), and may adapt to refusals.

Defender. The service provider controls (i) a retrieval index over O in V_R, (ii) an alignment mapping to V_M, and (iii) the inference-time UnRe procedure that computes the pre-check gate and updates $\tilde{v}_1 = v_1 + \delta$. The defender does *not* modify M's weights and performs *no training*. Gradients of M are available to the defender for the PGD updates; in strictly black-box deployments, zero-order variants can be used.

Security goal. For queries that hit O according to the pre-check gate, produce outputs that avoid targeted content while preserving on-topic semantics and minimizing side-effects on benign inputs.

E Additional Details of Methodology

E.1 Logit normalization and centering

For improved invariance and stability one may replace $z(t)$ by a normalized direction:

$$\textbf{Unit-only:} \quad \widehat{z}(t) \;=\; \frac{z(t)}{\|z(t)\|_2}. \tag{14}$$

The unit-only form is sufficient in practice; the centered variant adds invariance to constant bias shifts.

Time pooling We pool per-step directions and compare only the pooled vectors:

$$\bar{\widehat{z}}_\delta = \frac{1}{T_\delta}\sum_{t=1}^{T_\delta} \widehat{z}_\delta(t), \quad \bar{\widehat{z}}_0 = \frac{1}{T_0}\sum_{t=1}^{T_0} \widehat{z}_0(t), \quad R_{\text{dist}}^{\text{pool}} = \cos\big(\bar{\widehat{z}}_\delta,\, \bar{\widehat{z}}_0\big).$$

Use $R_{\text{dist}}^{\text{pool}}$ in Equation 15 as a drop-in replacement for R_{dist}.

E.1.1 Extension of Loss

$$\mathcal{L}(\delta) \;=\; \log\big(1 + \exp(\,\pi - S\,)\big) \tag{15}$$

$$\mathcal{L}(\delta) \;=\; \max(\pi - S,\, 0) \;+\; \log\big(1 + \exp\big(-|\pi - S|\big)\big) \tag{16}$$

Algorithm 2 UNRE Inference

Require: query q; LLM $\mathbf{M}$ with Embedding space $\mathrm{V_M}$; RAG embedding space V_R context v_1; optimized δ^J;

1: UNREed Matrix: $v_c \leftarrow v_1 + \delta^J$
2: **if** `EmbeddingAlignment` $= False$ **then**
3: UNREed Context: $C \leftarrow \mathrm{V_R}.\mathrm{Decode}(v_c)$
4: LLM Input: $I \leftarrow q \oplus C$
5: Matrix in LLM Embedding Space: $v_m \leftarrow \mathrm{V_M}(\phi(I))$
6: UNREed LLM output: $y_c \leftarrow \mathbf{M}.\mathcal{G}_{\mathrm{V_M}}(v_m)$
7: **else**
8: UNREed LLM output: $y_c \leftarrow \mathbf{M}.\mathcal{G}_{\mathrm{V_M}}(v_c)$
9: **end if**
10: *Restore* UNREed Context: C; UNREed Embedding Matrix: v_c
11: **return** UNREed LLM output: y_c

E.1.2 ADAPT LOSS TO DIFFERENT UNLEARNING TASKS

Following the loss design of Liu et al. (2024a), the UNRE Loss $\mathcal{L}$ can be extended into:

$$\mathcal{L}_{\text{unified}}(\delta) = \underbrace{\text{softplus}\big(\pi - S\big)}_{\text{UNRE base}} + \alpha \cdot \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \rho_a \, \ell_a\Big(s_a(\tilde{y}) \, , \, \omega_a\Big) \tag{17}$$

where,

- $\alpha \geq 0$ is a mixing coefficient that weights the ECO-style target term against the UNRE base term $\text{softplus}(\pi - S)$.
- $\mathcal{A}$ is the index set of task assessments (e.g., multiple-choice accuracy, BERTScore, ROUGE, ASG, PII hit-rate, etc.); $a \in \mathcal{A}$ indexes one assessment.
- $s_a(\tilde{y})$ denotes the score of the assessment a computed on the generated sequence $\tilde{y}$
- ω_a is the desired target for assessment a that represents a *retain-like* state in the sense of ECO Liu et al. (2024a).
- $\rho_a \geq 0$ is an optional weight for assessment a.
- $\ell_a(\cdot, \cdot)$ is a penalty for assessment a

E.1.3 TASK-SPECIFIC INSTANTIATION

We keep the loss form in Equation 17 unchanged and only plug in task-specific assessments $\langle \mathcal{A}, s_a, \omega_a, \ell_a, \rho_a \rangle$.

E.1.4 CONTEXT FORMING IN DIFFERENT SCENARIOS

Since there is a potential option that RAG can send the embedding vectors to LLM embeddings Cheng et al. (2024) , UNRE can align with the LLM so that direct connect the embedding space.

Based on the scenarios (when the LLM embeddings can not be aligned), UNREowner can decide if they want to deliver v_c to LLM Embedding $\mathrm{V_M}$ or decode into discrete natural language sentences and input M directly, thus skipping lines 2 to 7 in Algorithm 2.

Meanwhile, we store v_c and C, linked to v_1, as a UNRE-ed set that can be transferred to other unlearning methods for use as the *retain set*.

F ADDITIONAL EXPERIMENT SETTINGS

More introductions on baselines. GA: gradient-ascent on forget data to suppress target likelihood.

KL: GA with KL-to-reference regularization to preserve utility.

GD: gradient-based unlearning with direct loss on forget and a retain-side utility term (lightweight GA variant)

LLMU: a train-time unlearning recipe combining GA, random-mismatch loss, and KL-to-original for stability

PO: preference-style optimization that downranks forget-consistent responses relative to retain-consistent ones

DPO: direct preference optimization adapted to unlearning (no reward model)

NPO: negative preference optimization to avoid GA collapse and improve the forget/utility trade-off

FLAT: forget-data-only loss adjustment (no retain data / no reference model)

ICUL: in-context unlearning via specially constructed contexts and a likelihood-ratio signal at inference time

Prompt: rule-based *output filtering / guardrails* that refuse/deflect on forget-related queries

GUARD: detection + adaptive restriction during decoding to block forbidden tokens/semantics

ECO: embedding-corrupted prompts gated by a prompt classifier to enforce an "unlearned state" at inference

GAGDR/GAKLR: GA augmented with (i) gradient-direction regularization (GDR) or (ii) KL-to-retained anchoring (KLR) to stabilize utility (regularized GA variants).

NPOGDR/NPOKLR: NPO with the same (GDR/KLR) retain-side regularizers (regularized NPO variants).

Mismatch: context-mismatch baseline pairing queries with intentionally mismatched passages to reduce recall of copyrighted/entity text

G Additional Experiments

G.1 MUSE-News Unlearning

Experiment Setup We evaluate on **MUSE-News** with its two tasks: *VerbMem* (verbatim memorization) and *KnowMem* (knowledge memorization). Following the official protocol, we report four metrics: **VerbMem on** D_f (lower is better), **KnowMem on** D_f (lower is better), **KnowMem on** D_r (higher is better), and **PrivLeak** (closer to 0 is better; large positive/negative indicates leakage or under-unlearning).

It can be find that in **MUSE** benchmark in table. 5, UnRe can also overperform most of the baselines.

G.2 Copyright Harry Potter book - ECO (to make HP comparison with the main body)

For better comparison with **HP** benchmark, we put additional model performance in table. 6 tested on *qwen* model to extend the performance comparison.

G.3 Knowledge Unlearning WMDP

Table. 7 is showing the knowledge unlearning performance of GRUN, which can have better comparison with other methods.

G.4 Time Complexity

Components UnRe performs a one-time offline step to embed the forget set and build a lightweight index for the pre-check gate and then at inference runs a gated similarity check model

Table 5: MUSE-News results (official four metrics). Lower is better for VerbMem/KnowMem on D_f, higher is better for KnowMem on D_r, and PrivLeak should be close to 0. Rows from FLAT are complemented by additional training-free baselines and GUARD from its paper.

Method	**VerbMem on D_f ↓**	**KnowMem on D_f ↓**	**KnowMem on D_r ↑**	**PrivLeak**
Original LLM	58.4	63.9	55.2	−99.8
Retained LLM	20.8	33.1	55.0	0.0
Task Vectors	56.3	63.7	54.6	−99.8
WHP	19.7	21.2	28.3	109.6
GA	0.0	0.0	0.0	17.0
GD	4.9	27.5	6.7	109.4
KL	27.4	50.2	44.8	−96.1
NPO	0.0	0.0	0.0	15.0
NPO-RT	1.2	54.6	40.5	105.8
Mismatch	42.8	52.6	45.7	−99.8
FLAT (TV)	1.7	13.6	31.8	45.4
FLAT (KL)	0.0	0.0	0.0	58.9
FLAT (JS)	1.9	36.2	38.5	47.1
FLAT (Pearson)	1.6	0.0	0.2	26.8
ICUL	10.7	19.7	55.2	−99.8
Output Filtering	1.1	0.3	55.2	−99.8
Prompt	15.4	47.9	55.2	−99.6
GUARD	4.3	4.9	55.2	109.6
UnRe	4.0	33.2	55.2	−99.8

Table 6: HP unlearning under ECO's evaluation (**Qwen1.5-7B**). Lower ASG/PPL and higher Utility/Unique-Token are better.

Method	**ASG ↓**	**Utility ↑**	**PPL ↓**	**Unique(%) ↑**	**BERTScore**	**METEOR**	**ROUGE**	**SacreBLEU**
Original	30.9	51.1	1.3	56.9	83.9	50.1	43.8	34.0
Retain	0.0	53.2	1.9	13.2	65.8	11.7	9.4	1.1
Fine-tune	5.1	51.5	3.2	40.4	71.9	19.8	13.7	3.1
GA	8.6	33.2	–	0.4	53.4	0.0	0.0	0.1
GD	1.6	51.4	3.6	31.5	69.1	14.3	9.8	1.5
KL	6.9	51.6	2.2	38.3	73.4	23.0	14.7	4.6
Mismatch	5.1	51.8	3.0	39.7	71.8	19.6	13.8	3.2
SCRUB	9.4	31.5	–	4.3	48.4	1.0	0.8	0.2
LLMU	9.8	51.5	1.8	37.6	74.7	26.5	17.9	8.1
ECO	**2.8**	**51.1**	**1.8**	**29.1**	**57.6**	**11.8**	**6.6**	**1.3**

Table 7: WMDP results reported by GRUN. Bio/Cyber are accuracies (0–1). Chem was not reported.

Model	**Method**	**Bio ↓**	**Cyber ↓**	**MMLU ↑**
Llama 3.1	Before	0.696	0.418	0.611
	Vanilla	0.494	0.337	0.581
	GRUN	0.372	0.293	0.577
Mistral v0.1	Before	0.668	0.437	0.581
	Vanilla	0.256	0.252	0.529
	GRUN	0.293	0.278	0.535

weights are unchanged. Baselines like *LLMU* requires fine-tuning, and ICUL requires reverse-generation for context data.

Time Complexity (normalized)

It can be found in table. 8 the overall time complexity for lightweight scenario is the advantage of UnRe.

Table 8: **Time Complexity (normalized)**

Method	Offline Total	per-query	Overall for one run
LLMU	1.00	1.00	2.00
ICUL	0.11	1.15	1.26
UnRe (Ours)	0	1.67	1.67

Table 9: **TOFU 1% split**. Performance of our method and baseline methods on the TOFU dataset using two base LLMs (Llama2-7B and Phi-1.5B). FQ, MU, F-RL, and R-RL denote *forget quality*, *model utility*, *ROUGE-L on the forget set*, and *ROUGE-L on the retain set*, respectively. We include the Original LLM and the Retained LLM for reference. The top two results in each column are highlighted.

	Llama2-7B				Phi-1.5B			
Method	**FQ** ↑	**MU** ↑	**F-RL** ↓	**R-RL** ↑	**FQ** ↑	**MU** ↑	**F-RL** ↓	**R-RL** ↑
Original LLM	4.4883e-06	0.6346	0.9851	0.9833	0.0013	0.5184	0.9607	0.9199
Retained LLM	1.0	0.6267	0.4080	0.9833	1.0	0.5233	0.4272	0.9269
GA	0.0143	0.6333	0.4862	0.9008	0.0013	0.5069	0.5114	0.8048
KL	0.0068	0.6300	0.5281	0.9398	0.0030	0.5047	0.5059	0.8109
GradDiff	0.0068	0.6320	0.4773	0.8912	0.0030	0.5110	0.4996	0.8496
PO	0.0541	0.6308	0.3640	0.8811	0.0286	0.5127	0.3170	0.7468
Mismatch	0.0143	0.6304	0.9406	0.9741	0.0030	0.5225	0.9612	0.9194
LLMU	0.0030	0.5999	0.4891	0.9236	0.0143	0.5083	0.3380	0.7685
ICUL	0.0005	0.6239	0.4772	0.9818	0.0286	0.5195	0.0564	0.9276
Output Filtering	0.0002	0.6239	0.0	0.9818	0.00002	0.5195	0.0	0.9276
Prompt	0.0005	0.6239	0.5915	0.9818	0.0143	0.5195	0.1136	0.9276
DPO	0.0541	0.6359	0.5860	0.8852	0.0521	0.0519	0.3437	0.7349
NPO	0.0068	0.6321	0.4632	0.8950	0.0030	0.5057	0.5196	0.8000
FLAT (TV)	0.0541	0.6373	0.4391	0.8826	0.0143	0.5168	0.4689	0.8155
FLAT (KL)	0.0286	0.6393	0.5199	0.8750	0.0143	0.5180	0.4524	0.7850
FLAT (JS)	0.0541	0.6364	0.4454	0.8864	0.0068	0.5144	0.4572	0.8117
FLAT (Pearson)	0.0541	0.6374	0.4392	0.8857	0.0143	0.5175	0.4591	0.8099
ECO (Rand Noise)	0.9188	0.6257	0.0538	0.9798	0.7659	0.5519	0.2310	0.9213
ECO (Zero-Out)	0.9900	0.6257	0.5182	0.9798	0.9900	0.5519	0.4143	0.9213
GUARD	0.1649	0.6239	0.3910	0.9818	0.1649	0.5195	0.4214	0.9276
UnRe (Ours)	**0.8087**	**0.6259**	**0.3297**	**0.9976**	**0.7566**	**0.5117**	**0.2176**	**0.9321**

G.5 UNLEARNING PERFORMANCE COMPARING WITH RAG

Since the traditional RAG-based unlearning methods can just reduce the RAG augmenting performance through reranking or unlearning through loading retain document Wang et al. (2024b), or require a *retain set*, thus UNRE is not comparable with traditional RAG-based unlearning methods.

G.6 TOFU 1% SPLIT ON MORE MODELS

UNRE preserves model utility. As shown in Table 9, UNREincurs almost no degradation in model utility compared to the original/retained references. On **Llama2-7B**, UnRe attains a top–2 MU, on par with the best FLAT variant. On **Phi-1.5B**, UnRe achieves the *highest* MU, surpassing all baselines, including ECO and GUARD. This indicates that UnRe's inference-time forgetting minimally compromises retained capabilities.

UNRE delivers top-tier Forget Quality. UnRe attains very strong FQ on both LLMs, ranking among the top results. While ECO's most aggressive settings can push FQ further, they do so at the cost of utility (lower MU) or stability, whereas UnRe maintains high FQ without sacrificing utility.

UNRE achieves a better trade-off between forgetting and retention. UNRE substantially reduces **F-RL** (forget-side ROUGE-L) on Llama2-7B and on Phi-1.5B—while keeping **R-RL** (retain-side ROUGE-L) near the top. Compared with FLAT and GUARD, UNRE consistently attains stronger

forgetting (lower F-RL, higher FQ) *and* stronger utility/retention (higher MU and R-RL), yielding the most favorable balance overall on both model families.

G.7 HAZARDOUS KNOWLEDGE UNLEARNING

We evaluate hazardous-knowledge unlearning on **WMDP** (Bio/Chem/Cyber; 4-choice MCQ) following the ECO protocol: we report per-domain **MCQ accuracy on the *forget set*** ($\downarrow$ is better; random guess is 25%) as the unlearning signal, together with **MMLU** accuracy ($\uparrow$ is better) as a model-utility proxy on the retain/general side. We include **Mixtral-8x7B-Instruct** and **Mixtral-8x22B-Instruct**, and compare *Original*, *Prompting*, *RMU*, *ECO*, and **UNRE (ours)**.

Table 10: **WMDP hazardous-knowledge unlearning**. Bio/Chem/Cyber are multiple-choice accuracies on the forget set ($\downarrow$), and MMLU is utility on the retain side ($\uparrow$).

Model	Method	Bio $\downarrow$	Chem $\downarrow$	Cyber $\downarrow$	MMLU $\uparrow$
Mixtral-8x7B-Instruct					
	Original	71.6	53.4	51.9	67.7
	Prompting	46.4	37.0	47.7	61.9
	RMU	32.0	52.7	31.4	66.1
	ECO	25.0	23.4	26.4	67.7
	UnRE (Ours)	**29.2** $\pm$ 1.1	**49.6** $\pm$ 1.5	**30.3** $\pm$ 1.2	**65.2** $\pm$ 0.7
Mixtral-8x22B-Instruct					
	Original	77.3	56.6	52.6	73.9
	Prompting	56.4	45.6	42.5	69.8
	ECO	26.7	23.9	24.1	73.9
	UnRE (Ours)	**26.3** $\pm$ 0.6	**19.6** $\pm$ 1.0	**17.7** $\pm$ 1.3	**69.7** $\pm$ 0.4
Random guess		25.0	25.0	25.0	25.0

As shown in table 10, the results align with the intended behavior of inference-time unlearning, UNRE provides the best balance on the larger model.

H VISUALIZATION

We visualize some of the examples with selected tokens and pieces so as to vividly express the effectiveness of UNRE. Details can be seen in Table 11.

Table 11: **UNRE dynamic context examples (5 queries $\times$ 3 contexts).** Pre = retrieved forget context; Post = UnRe-modified embedding decoded contexts (selected).

Query	Ctx	Retrieved forget context (Pre)	UNRE modified context (Post)
Q1: Where was Bruce Lee born?	C1	Bruce Lee was born in San Francisco.	Birthplace: U.S. city (withheld).
	C2	He was born in San Francisco's Chinatown.	Neighborhood withheld; city unspecified.
	C3	He was born in 1940.	Birth year: early 1940s (approx.).
Q2: Which company is Warren Buffett chairman and CEO of?	C1	He is chairman and CEO of Berkshire Hathaway.	He leads a U.S. conglomerate (name withheld).
	C2	The firm is headquartered in Omaha, Nebraska.	Headquarters: Midwestern U.S. (city masked).
	C3	He is nicknamed the Oracle of Omaha.	Nickname tied to hometown; redacted.
Q3: What is the philosophy taught by Confucius called?	C1	The philosophy is known as Confucianism.	Name of school withheld.
	C2	Confucius's family name was Kong.	Family name omitted.
	C3	Zhong You was one of his disciples.	Disciple name withheld.
Q4: Which song did Christina Aguilera record for the 1998 film *Mulan*?	C1	She recorded the theme song "Reflection" for *Mulan*.	Theme song title withheld.
	C2	She won the Grammy for Best New Artist in 2000.	Major "Best New Artist" award circa 2000.
	C3	She was later named a Disney Legend.	Received Disney lifetime honor.
Q5: What was Stephen King's debut novel?	C1	Stephen King's debut novel was "Carrie".	Debut novel title withheld.
	C2	He is widely called the "King of Horror".	Strongly linked to the horror genre.
	C3	In 1982 he published "Different Seasons".	Published a novella collection (early 1980s).