Learning to Erase Private Knowledge from Multi-Documents for Retrieval-Augmented Large Language Models

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) is a promising technique for applying LLMs to proprietary domains. However, retrieved documents may contain sensitive knowledge, posing risks of privacy leakage in generative results. Thus, effectively erasing private information from retrieved documents is a key challenge for RAG. Unlike traditional text anonymization, RAG should consider: (1) the inherent multi-document reasoning may face de-anonymization attacks; (2) private knowledge varies by scenarios, so users should be allowed to customize which information to erase; (3) preserving sufficient publicly available knowledge for generation tasks. This paper introduces the privacy erasure task for RAG and proposes Eraser4RAG, a private knowledge eraser that effectively removes user-defined private knowledge from documents while preserving sufficient public knowledge for generation. Specifically, we first construct a global knowledge graph to identify potential knowledge across documents, aiming to defend against de-anonymization attacks. Then we randomly split it into private and public sub-graphs, and fine-tune Flan-T5 to rewrite the retrieved documents, excluding private triples. Finally, PPO algorithm optimizes the rewriting model to minimize private triples and maximize public triples retention. Experiments on four QA datasets demonstrate that Eraser4RAG achieves superior erase performance than GPT-4o.

1 Introduction

Retrieval-augmented generation (RAG) is regarded a promising solution to deploy large language models (LLMs) in proprietary domains (Fan et al., 2024; Zeng et al., 2024a,b), as it leverages proprietary data as a retrieval corpus to obtain relevant documents for generation.

However, documents retrieved from proprietary data may contain sensitive information, such as home addresses or social relationships, posing risks of privacy leakage in generative results (Zeng et al., 2024a). Therefore, it is crucial to design an efficient private knowledge eraser for RAG, capable of removing sensitive information from retrieval documents while preserving public knowledge to support downstream generation. Traditional text anonymization methods (Dou et al., 2024a; Frikha et al.; Mosallanezhad et al., 2019) rely on obfuscating expressions or adding noise to erase specific private knowledge in a document. They leverage differential privacy techniques that add noise during training and tokenization to remove private information, or use synthetic data generation, document sanitization, and privacy-aware retrieval to preserve utility while minimizing exposure of private information.

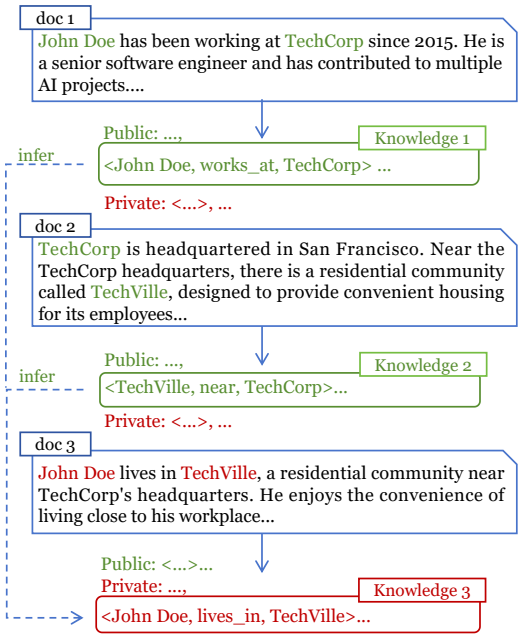


Figure 1: Example of de-anonymization through multiple documents reasoning.

Unlike traditional text anonymization tasks, the **Privacy Erasure** task in RAG scenarios presents more complex challenges: (1) The combination

of multiple documents can pose a risk of de-anonymization attacks (Ji et al., 2016). Even if a single document does not contain private information, aggregating multiple documents can still lead to the inference of private knowledge (Hoang et al., 2023). As shown in Figure 1, doc1 and doc2 each contain public knowledge triples, Knowledge1 and Knowledge2. But these public triples together can be used to infer Knowledge3, which involves private information about "the living address of John Doe". (2) The definition of private knowledge varies across different scenarios, e.g., a celebrity's birthday is public, but a teenager's is not. Privacy information not only includes standalone non-relational private entities (such as Personally Identifiable Information, PII), but also fine-grained relational knowledge across the retrieved documents. Therefore, a human-centric RAG eraser should allow precise, controllable removal based on user definition. (3) Retaining adequate public knowledge is crucial for generative model reasoning, to ensure RAG remains practical and valuable.

In this paper, we propose **Eraser4RAG**, a private knowledge Eraser that rewrites documents to maximize public knowledge while minimizing user-defined private knowledge. Specifically, we first extract knowledge triples and entities from each retrieved document and construct a global knowledge graph to identify potential private knowledge across documents, reducing de-anonymization risks. Then, a connected subset of the graph is selected as the private graph, and the remainder forms the public graph. Given the original documents, private graph, and public graph, we fine-tune Flan-T5 model to generate rewritten documents excluding private triples and entities. Finally, we explore the optimal rewriting strategy for Eraser4RAG using PPO algorithm to maximize the retention of public knowledge and minimize the presence of private knowledge. During inference, the private knowledge triples or entities are firstly provided by the upstream data controller according to domain- or user-specific requirement. And then, our rewriting model focuses on removing the annotated private content from the retrieved documents while retaining valuable public knowledge for downstream QA tasks.

Experimental results on four public QA datasets demonstrate that Eraser4RAG achieves better erase performance than GPT-4o, with the public triple ratio increasing by 1.8% and the private triple ratio decreasing by 21.8%. Furthermore, Eraser4RAG has a smaller impact on downstream generation performance than baselines, validating the effectiveness of our Eraser4RAG[1].

The innovations of this paper are as follows:

- In RAG scenario, we introduce the privacy erasure task, which requires addressing the issues of multi-document de-anonymization, controllable erasure, and available reasoning.
- We propose Eraser4RAG, the first model for controlled private knowledge erasure for RAG, reducing private knowledge while retaining essential public knowledge through RL.
- Experimental results show that Eraser4RAG effectively maximizes public triples retention and minimizes private triples, validating the effectiveness of our RL-based strategy.

2 Privacy Erasure Task

In RAG, the retrieved documents $D = \{d_i\}_{i=1}^k$ may contain private information that needs to be erased before further LLM reasoning. Let $I_{\text{pri}}(D)$ and $I_{\text{pub}}(D)$ denote the private and public information in D, respectively. The privacy erasure model is denoted as $\mathcal{M}_\theta$, where θ represents the model parameters. The processed document set is given by $D' = \{d'_i\}_{i=1}^k = \{\mathcal{M}_\theta(d_i)\}_{i=1}^k$. Then the objective of privacy erasure is:

$$\min_\theta \left[|I_{\text{pri}}(D')|,\ -|I_{\text{pub}}(D')|\right].$$

In RAG scenario, the definitions of private and public information, I_{pri} and I_{pub}, are flexible and random. Due to the challenge of cross-document de-anonymization, it is insufficient to minimize the private information $I_{\text{pri}}(d'_i)$ of individual documents alone; instead, the private information of the entire document set must be considered.

We use knowledge graph triples as the representation unit to quantify erasure effectiveness, since it enables: (1) fine-grained control over the removal of private information by precisely locating and manipulating specific knowledge units; and (2) quantitative evaluation of privacy erasure and public information preservation through triple-level comparison. Such fine-grained triples enable targeted deletion and accurate assessment of erasure quality.

Both private and public information can be expressed as entity-relation triples $[h, r, t]$, where the

[1]Our code and annotated dataset are available at: https://anonymous.4open.science/r/Eraser4RAG-4E39/

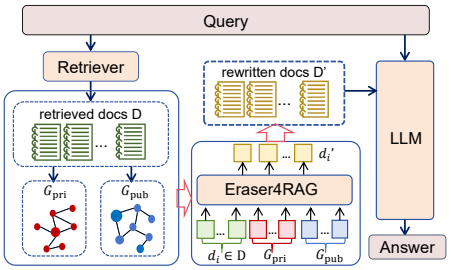


Figure 2: RAG pipeline with privacy erasure.

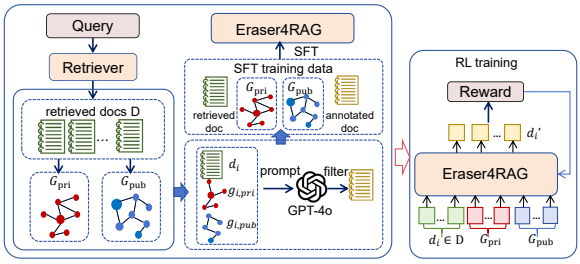


Figure 3: The training process of Eraser4RAG.

head entity h and the tail entity t represent the subject and object of the relation r, respectively. In addition, for a non-relational private entity A to be deleted (e.g., only PIIs, temporal dimensions), we represent it in the form of a self-loop triple, that is, $[A, A, A]$. This enables us to uniformly handle both relational and non-relational private knowledge within a triple-based framework. Private and public information within D are thus represented as triple sets G_{pri} and G_{pub}. The partitioning of private and public information must satisfy: Neither private information can be inferred from any combination of public information, nor public information from any combination of private information.

Given this setup, our task is to train a privacy-aware rewriting model $\mathcal{M}_\theta$ to transform each document d_i into its rewritten version d'_i, ensuring that private information is removed while public information is retained, $d'_i = \mathcal{M}_\theta(d_i, G_{\text{pri}}, G_{\text{pub}})$.

For privacy-preserved RAG service,we assume the data controller (e.g., banks or hospitals) defines and annotates private information in accordance with domain-specific regulations. The pipeline is shown in Figure 2. Given a query q, the retriever first retrieves documents $D = \{d_i\}_{i=1}^k$. The data controller identifies the private triples G_{pri} and public triples G_{pub} from documents. The rewriting model then rewrites each $d_i \in D$ according to G_{pri} and G_{pub}, producing the rewritten documents $D' = \{d'_i\}_{i=1}^k$. Given the query q and the rewritten documents D', LLM generates an answer, $a = \text{LLM}(q, D')$.

3 Approach

The training process of Eraser4RAG consists of two stages: supervised fine-tuning and reinforcement learning (RL), as shown in Figure 3. First, GPT-4o generates rewritten versions of the documents, from which high-quality rewrites are selected. Next, these selected rewrites are used to supervise fine-tuning Flan-T5 model, enabling it to acquire an initial rewriting capability. Finally, the fine-tuned model undergoes multi-objective RL, optimizing it to remove private information while preserving as much public information as possible.

3.1 Supervised Fine-tuning

This section outlines the annotated data for training, including privacy definitions, GPT rewrites, and data filtering, followed by the training method.

3.1.1 Rewritten Data Annotation

Definition of Private Information To train a model capable of removing private information from documents, it is necessary to define what constitutes private information in the training data. Since the definition of private information is subjective and varies across RAG service providers, we do not designate entity types or relations as private in our training data. Instead, we adopt a random selection strategy to simulate diverse privacy definition scenarios.

Specifically, given a query q and top-k relevant documents $D = \{d_i\}_{i=1}^k$, a relation extraction model extracts entity-relation triples $g_i = \{[h, r, t]\}$ from the document d_i. These extracted triples from all retrieved documents are merged into a knowledge graph G, representing a global set of all triples in documents D. A random subset of triples from G is selected as the candidate private set G_{pri}[2]. The remaining triples in G, after excluding those designated as private, form the candidate public information set G_{pub}. To ensure a clean separation between private and public information(see Section 2), we apply the following filtering steps: (1) For G_{pub}, delete triples whose head and tail entities are connected in the private G_{pri}, since these triples can be inferred with private information. (2) For G_{pri}, delete triples whose head and tail entities are connected in the public G_{pub}, since these triples can be inferred from G_{pub}.

[2]It is worth noting that the triples appearing in the query are not considered private information.

Rewritten by GPT To annotate the training data for the supervised fine-tuning (SFT) stage, we prompt GPT-4o to rewrite a given document based on the provided private and public information sets to remove all private information, retain all public information, and maintain the language style.

Specifically, the private information provided to GPT-4o is the intersection of the entity-relation triples set g_i extracted from document d_i with the global private knowledge set G_{pri}, resulting in $g_{i,\text{pri}} = g_i \cap G_{\text{pri}}$. The public triples are defined similarly as $g_{i,\text{pub}} = g_i \cap G_{\text{pub}}$. The prompt used for data annotation is shown in Figure 5. Note that the 'private' and 'public' triples are extracted from public datasets, so no privacy is transmitted to API.

Data Filtering To ensure the quality of the training data, we filter the GPT-4o rewritten documents. Specifically, we apply a relation extraction model to each rewritten document d'_i to extract the set of entity-relation triples g'_i. We compute the intersections of g'_i with the original document's local private set $g_{i,\text{pri}}$ and public set $g_{i,\text{pub}}$, obtaining the private set $g'_{i,\text{pri}}$ and the public set $g'_{i,\text{pub}}$ for d'_i.

To quantitatively evaluate the quality of the rewritten documents, we define two retention rates: privacy retention rate $r_{\text{pri}} = \frac{\#g'_{i,\text{pri}}}{\#g_{i,\text{pri}}}$, which measures the proportion of private information retained, and public information retention rate $r_{\text{pub}} = \frac{\#g'_{i,\text{pub}}}{\#g_{i,\text{pub}}}$, proportion of public information retained. Rewritten documents with $r_{\text{pri}} = 0$ and $r_{\text{pub}} > 0.8$ are selected into the SFT dataset D_c.

3.1.2 Training Rewriting Model

We fine-tune the pretrained Flan-T5 model on the annotated dataset to equip it with basic privacy removal capabilities. Following a similar approach to Huguet Cabot and Navigli (2021), we introduce special tokens to linearize the private and public triple sets, ensuring compatibility with the language model's input format. A public triple $[h_{\text{pub}}, r_{\text{pub}}, t_{\text{pub}}]$ is represented as:

$$\texttt{<csubj>}h_{\text{pub}}\texttt{<crel>}r_{\text{pub}}\texttt{<cobj>}t_{\text{pub}}\texttt{<ce>},$$

and a private triple $[h_{\text{pri}}, r_{\text{pri}}, t_{\text{pri}}]$ is represented as:

$$\texttt{<rsubj>}h_{\text{pri}}\texttt{<rrel>}r_{\text{pri}}\texttt{<robj>}t_{\text{pri}}\texttt{<re>}.$$

The original document d, the linearized global private triples G_{pri} and public triples G_{pub} are concatenated to form the input to the rewriter, allowing it to learn to identify all relevant triples related to the document content from the global set, rather than solely processing the privacy information defined within a single document. This design enables the model to handle cross-document de-anonymization challenges effectively.

Given an original document d, we denote $d_+ \in D_c$ as the rewritten version produced by GPT. The objective of supervised fine-tuning is to optimize the parameters θ of the rewriter $\mathcal{M}_\theta$ by minimizing the cross-entropy loss between the model's prediction d' and the target d_+:

$$L_{sft} = -\sum_{t=1}^{T} y_{d+}^{(t)} \log y_{d'}^{(t)},$$

where T is the number of tokens, $y_{d+}^{(t)}$ is the one-hot vector for the t-th token in the target sequence, and $y_{d'}^{(t)}$ is the predicted token probability distribution.

3.2 Reinforcement Learning

Although the rewriter acquires a basic ability to remove relevant private information after supervised fine-tuning, further optimization is required to retain sufficient essential public information while removing as much private information as possible. We employ RL framework with PPO algorithm (Schulman et al., 2017; Ziegler et al., 2020) to further optimize Eraser4RAG.

The reward function R is designed based on the privacy retention rate r_{pri} and the public information retention rate r_{pub}, evaluating the rewriting:

$$R = r_{\text{pub}} \cdot \exp(-p \cdot r_{\text{pri}}),$$

where r_{pub}, $r_{\text{pri}} \in [0, 1]$, and p is a dynamically adjusted parameter. R balances two competing objectives: maximizing the retention of public information while minimizing the presence of private information. The parameter p controls the privacy–utility trade-off.

4 Experiments

4.1 Tasks and Evaluation Metrics

Regarding the challenges we introduce the experimental tasks and evaluation metrics.

4.1.1 Privacy Removal and Public Knowledge Preservation

To evaluate the erasing performance, we select the specialized test set D_{special} from the retrieved documents set D:

$$\begin{aligned} D_{\text{special}} = \{ & d_i \in D \mid \exists\, [h, r, t]_{\text{pub}} \in g_{i,\text{pub}}, \\ & [h, r, t]_{\text{pri}} \in g_{i,\text{pri}}, \text{s.t. } t_{\text{pub}} = t_{\text{pri}} \}. \end{aligned} \tag{1}$$

Model	PopQA	TQA	NQ	HQA
Abstraction	0.1551	0.1808	0.1489	0.1442
Anonymizer	0.2463	0.2846	0.2765	0.2593
IncogniText	0.2865	0.1626	0.1522	0.1787
NER	**0.0180**	**0.0399**	**0.0493**	**0.0176**
gpt-each	0.1753	0.1613	0.1884	0.1696
gpt-all	0.3069	0.2467	0.2730	0.2767
Eraser4RAG	0.1099	0.1463	0.1429	0.1436

(a) Comparison of r_{pri} $\downarrow$ across different baselines on $D_{special}$ test set of PopQA, TQA, NQ and HQA.

Model	PopQA	TQA	NQ	HQA
Abstraction	0.4384	0.4247	0.3997	0.4107
Anonymizer	0.2386	0.2819	0.2684	0.2477
IncogniText	0.4553	0.3140	0.2927	0.3477
NER	0.0193	0.0427	0.0471	0.0193
gpt-each	**0.4825**	0.4337	0.4512	0.4628
gpt-all	0.4040	0.3323	0.3632	0.3790
Eraser4RAG	0.4793	**0.4441**	**0.4663**	**0.4720**

(b) Comparison of r_{pub} $\uparrow$ across different baselines on $D_{special}$ test set of PopQA, TQA, NQ and HQA.

Table 1: Performance comparison of different models in private and public information retention.

Model	Abstraction	Anonymizer	IncogniText
$r_{\text{remove}}(\%)$	100	71.6	80.4
PPL	1.35	1.29	1.25
Model	**NER**	**GPT-4o**	**Eraser4RAG**
$r_{\text{remove}}(\%)$	99.8	84.7	93.7
PPL	1.26	1.21	1.23

Table 2: Performance comparison of different models for private entities in r_{remove} $\uparrow$ and PPL $\downarrow$.

D_{special} consists of documents where public and private information share the same tail entity but different head entity or relation. Experiments on D_{special} mimic a scenario to remove private relationships while retaining public information, for fine-grained evaluation of anonymization.

We employ private information retention rate r_{pri} and public information retention rate r_{pub} as assessing metrics:

$$r_{\text{pri}} = \#g'_{i,\text{pri}}/\#g_{i,\text{pri}};\ r_{\text{pub}} = \#g'_{i,\text{pub}}/\#g_{i,\text{pub}},$$

where $\#g'_{i,\cdot}$ is the number of triples remaining in the rewritten document d'_i, and $\#g_{i,\cdot}$ is the original count of triples in d_i. A lower r_{pri} indicates better privacy removal, and a higher r_{pub} signifies better public information preservation.

In addition, we consider a standalone privacy removal task for entities that cannot be expressed as triples. Anonymization is successful if the entity no longer appears in the rewritten document. We define the entity removal rate as $r_{\text{remove}} = N_{\text{remove}}/N_{\text{total}}$, where N_{remove} is the number of documents which private entities are successfully removed, and N_{total} is the total number of documents to be anonymized. Fluency of rewritten texts is measured using perplexity (PPL) (Jelinek et al., 1977), where lower PPL means higher fluency.

Model	PopQA	TQA	NQ	HQA
Abstraction	0.3728	0.4069	0.4059	0.4135
Anonymizer	0.2996	0.3024	0.4411	0.3698
IncogniText	0.3923	0.4002	0.3959	0.4284
NER	**0.0221**	**0.0293**	**0.0341**	**0.0291**
gpt-each	0.3665	0.3811	0.3749	0.3775
gpt-all	0.4352	0.4413	0.4366	0.4407
Eraser4RAG	0.2600	0.3316	0.2102	0.2821

Table 3: Performance comparison of different models in r_{connect} $\downarrow$ on $\{D\}_{\text{infer}}$ test set.

4.1.2 Resistance to De-Anonymization

To assess whether the erasing method can resist cross-document de-anonymization, we construct a retrieval document subset that is susceptible to such attacks. The inference risk subset is:

$$\{D\}_{\text{infer}} = \{D \mid \exists\, d_i \in D,\ (h,t) \in g_{i,\text{pri}} \quad s.t.\ h \not\sim_{G_{\text{pri}} \setminus g_{i,\text{pri}}} t \wedge h \sim_{G_{\text{unpri}} \setminus g_{i,\text{unpri}}} t\}, \tag{2}$$

where $G_{\text{unpri}} = G \setminus G_{\text{pri}}$ is global unprivate knowledge graph of documents set D, and $g_{i,\text{unpri}} = g_i \setminus g_{i,\text{pri}}$ is unprivate triple set of individual document d_i. $G_{\text{unpri}} \setminus g_{i,\text{unpri}}$ represents the unprivate knowledge graph of other documents $D \setminus \{d_i\}$. It simulates cases where privacy is removed from one document but remains inferable through non-private information in other documents.

We evaluate de-anonymization risk on $\{D\}_{\text{infer}}$ by constructing an entity-relation graph G' based on the anonymized document set D'. For each privacy-sensitive triple $[h,r,t]_{\text{pri}}$ in the original private triple set G_{pri}, if h_{pri} and t_{pri} remain connected in G' via intermediate entities and relations from other documents, we consider the private information $[h,r,t]_{\text{pri}}$ to be potentially inferable. Therefore, we measure the risk of de-anonymization using the privacy connection ratio:

$$r_{\text{connect}} = \#\{[h,r,t]_{\text{pri}} \in G_{\text{pri}} \mid h \sim_{G'} t\}/\#G_{\text{pri}}.$$

This evaluation inherently accounts for multi-hop reasoning. A higher r_{connect} indicates a higher risk due to multi-document reasoning.

4.1.3 Utility on Downstream Task

To evaluate the utility of anonymized documents, we assess their performance on downstream tasks

using four QA datasets. Specifically, we measure the accuracy (acc) when the question and answer do not involve private information, that is, when the entities in the question and answer are not connected in the private knowledge graph G_{pri}.

4.1.4 Resistance to Privacy Attacks

Previous work (Zeng et al., 2024a) evaluates privacy leakage by prompting LLMs to reproduce private content (e.g., phone numbers, specific sentences). However, our scenario involves multi-document contexts where private knowledge often emerges through entity–relation associations across documents, rather than appearing in a single sentence. We design a QA-based privacy attack, where question–answer pairs reflect private knowledge. We first select those queries that LLM can answer before anonymization, and construct the privacy triples accordingly. During evaluation, anonymized documents are provided, and the accuracy of these queries defines the attack success rate r_{attack}. Lower r_{attack} indicates stronger resistance to inference-based privacy attacks.

4.2 Experimental Settings

4.2.1 Datasets

We conduct model training and evaluation on four QA datasets: PopQA (Mallen et al., 2023), TriviaQA (Joshi et al., 2017), Natural Questions(NQ) (Kwiatkowski et al., 2019), and HotpotQA (Yang et al., 2018), retrieving relevant documents from the Wikipedia corpus. Additionally, we validate the model's transferability on the NewsQA (Trischler et al., 2017) dataset. Details are provided in Appendix B.

4.2.2 Baselines

To validate the effectiveness of Eraser4RAG, we compare it with six baseline methods. The self-disclosure abstraction model (Abstraction) (Dou et al., 2024b) rephrases sensitive self-disclosure spans into less specific terms, generating diverse and privacy-preserving alternatives. Anonymizer (Staab et al., 2024) proposes a feedback-guided adversarial anonymization framework to iteratively anonymize text. IncogniText (Frikha et al., 2024) rewrites text with incorrect private attribute values to mislead adversaries. NER (Asimopoulos et al., 2024) identifies and replaces all entities in the document. Additionally, we employ two GPT-4o (Hurst et al., 2024) rewriting strategies: gpt-all using the global private triple set G_{pri} and public triple set G_{pub} as prompts, and gpt-each utilizing only the private triple set $g_{i,\text{pri}}$ and public triple set $g_{i,\text{pub}}$ from the individual retrieved document d_i. The GPT-4o prompt is provided in Figure 5.

4.2.3 Implementation Details

We use Llama3-8b-instruct as the base LM for the generator and Flan-T5-large (Chung et al., 2024) as the base model for privacy-preserving text rewriting. Our rewriting model only has 770M parameters, which can run locally in resource-constrained settings. For retrieval, we adopt the off-the-shelf Contriever-MS MARCO model (Izacard et al., 2022) by default, retrieving up to ten documents for each input. The relation extraction model is relik-relation-extraction-small (Orlando et al., 2024). Parameter details are provided in Appendix A.

4.3 Main Results

4.3.1 Private Information Erasure

Table 1a compares the private information retention rate r_{pri} across models on $D_{special}$ test set from four QA datasets. Eraser4RAG consistently achieves the lowest r_{pri} among baselines except for NER. Among the baselines, gpt-all exhibits the highest retention rate, suggesting that using a global privacy-aware prompt for GPT-4o is less effective in eliminating sensitive information compared to document-specific rewriting. Abstraction model remains less effective than Eraser4RAG due to replacing private spans is unable to fully obscure sensitive relations.

Table 2 compares standalone private entity removal. Eraser4RAG achieves higher r_{remove} than baselines with better fluency, indicating that the self-loop triple formulation effectively supports the removal of standalone entities. Abstraction method attains 100% removal rate by direct span replacement, but leads to ungrammatical output and lower fluency. GPT-4o shows high fluency with 200B parameters but a lower removal rate than Eraser4RAG with 770M parameters, confirming Eraser4RAG's consistent handling of both relational and standalone privacy.

4.3.2 Public Information Preservation

Table 1b compares r_{pub} across models. Eraser4RAG achieves the highest performance on three out of the four datasets (TQA, NQ, and HQA). gpt-each performs slightly better than Eraser4RAG on PopQA but underperforms on the other datasets, suggesting limited consistency

Model	PopQA	TQA	NQ	HQA
w/o rewrite	0.4292	0.6655	0.3865	0.2690
Abstraction	0.3785	0.6550	**0.3635**	0.2576
Anonymizer	0.2252	0.5655	0.2179	0.2029
IncogniText	0.3824	0.5986	0.3423	0.2218
NER	0.1992	0.1438	0.1988	0.1284
gpt-each	**0.3919**	0.6484	0.3460	0.2613
gpt-all	0.3868	0.6443	0.3456	**0.2640**
Eraser4RAG	0.3870	**0.6635**	0.3568	0.2585

Table 4: RAG accuracy↑ comparison of different methods' rewritting across four datasets.

when using per-document triples. gpt-all, relying on global privacy constraints, achieves the lowest scores due to excessive information removal. Abstraction model fails to outperform Eraser4RAG or gpt-each, as generalizing sensitive spans into less specific terms without considering the relation between entities leads to more information loss.

4.3.3 Resistance to De-Anonymization

Table 3 compares r_{connect} across different methods on the corresponding test set D_{connect} of four datasets. Except NER, Eraser4RAG consistently achieves the lowest r_{connect} across datasets, demonstrating effective prevention of private entity pairs from being reconnected in the anonymized graph. In contrast, gpt-all exhibits the highest r_{connect}, suggesting that aggregating all private and public triples for rewriting results in excessive information interference. Abstraction model shows comparable performance to gpt-each, suggesting that replacing sensitive spans with abstracted terms does not fully break cross-document linkability.

4.3.4 Downstream RAG Accuracy

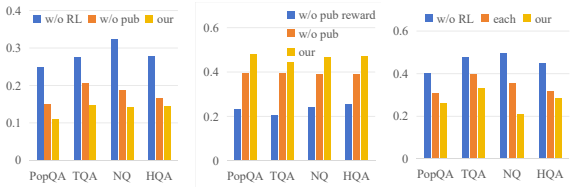
Table 4 compares RAG accuracy using rewritten documents. Eraser4RAG maintains competitive performance despite eliminating private relationships. While some answer-relevant connections may be lost due to privacy removal, the model still preserves enough useful information to maintain RAG effectiveness. On the multi-hop dataset HQA, where answers rely on cross-document reasoning, Eraser4RAG's rewriting tends to remove relations that are crucial for deriving the final answer, leading to a larger performance drop.

4.3.5 Resistance to Privacy Attacks

Table 5 compares privacy leakage in the QA-based attack setting. From the results, Eraser4RAG obtains lower r_{attack} than most baselines, indicating the effectiveness of privacy protection. Compared

Model	Abstraction	Anonymizer	IncogniText
r_{attack} ↓	0.4001	0.3788	0.4375
Model	**gpt-each**	**gpt-all**	**Our**
r_{attack} ↓	0.4338	0.4252	0.3745

Table 5: Evaluation of privacy leakage under QA-based attack setting.



(a) r_{pri} ↓ (b) r_{pub} ↑ (c) $r_{connect}$ ↓

Figure 4: Ablation study results on different datasets for metrics: (a) r_{pri}, (b) r_{pub}, and (c) r_{connect}. The comparisons include variants without RL (w/o RL), without public triplets (w/o pub), without reward related to public information (w/o pub reward), only inputting triplets from the individual document (each), and our full method (our).

with other methods, Eraser4RAG reduces the leakage level by using a privacy erasure approach that targets the relationships between entities, indicating its strong resistance to defend against inference-based privacy attacks.

4.4 Ablation Study

4.4.1 Public Information Constraint

To verify the effectiveness of public triples during rewriting, we evaluate when only the original document and the private triple set are provided as input (w/o pub). The results, shown in Figures 4a and 4b, demonstrate the impact on r_{pri} and r_{pub}. Due to the disjoint nature of public and private information, incorporating public triples when rewriting not only constrains the rewriting from arbitrarily removing content unrelated to private information, but also provides a comparative signal that helps the rewriter better identify privacy.

4.4.2 RL Training

To verify the effectiveness of RL, we evaluate the privacy anonymization capability of the rewriter trained only with SFT, as shown in Figure 4a and 4c. Due to limitations in the quantity and quality of annotated data used for SFT, a rewriter trained solely with SFT fails to effectively remove privacy-sensitive relations and prevent cross-document de-anonymization. This further highlights that RL

model	$r_{\text{pub}} \uparrow$	$r_{\text{pri}} \downarrow$	$r_{\text{connect}} \downarrow$
Abstraction	0.3332	**0.1075**	0.4345
gpt-each	0.4272	0.1636	0.4150
gpt-all	0.3321	0.2558	0.4584
Eraser4RAG	**0.4355**	0.1124	**0.2622**

Table 6: Erase performance on NewsQA.

can guide the model toward the desired objective through reward feedback.

4.4.3 Multi-objective Reward

To verify the necessity of multi-objective rewards in RL, we evaluate the rewriter's ability when using only the privacy-related reward $R = \exp(-p \cdot r_{\text{pri}})$, as shown in Figure 4b. It can be observed that omitting the constraint on r_{pub} in the reward function leads to excessive removal of non-private content, reducing the model's ability to preserve relevant public information and potentially impacting downstream task performance.

4.4.4 Global Triple Set

To validate that using the global triple sets G_{pri} and G_{pub} helps mitigate cross-document de-anonymization, we evaluate the performance when prompting the rewriter with only the triples $g_{i,pri}$ and $g_{i,pub}$ relevant to the single document d_i, as shown in Figure 4c. The results indicate that relying solely on document-specific triples significantly increases the risk of cross-document de-anonymization. Therefore, providing the global triple set G_{pri} imposes stronger constraints on the rewriter, ensuring the elimination of relations that could infer private information beyond those explicitly present within the given document.

4.5 Transferability

To verify the transferability of Eraser4RAG trained on the Wikipedia corpus, we evaluate its performance on NewsQA dataset, which consists of CNN news articles. Although Abstraction model achieves the lowest private information retention rate, it struggles to maintain public knowledge, resulting in a lower r_{pub}. Eraser4RAG can generalize well to different textual domains, accurately identifying and eliminating privacy-related connections while preserving essential public information.

5 Related Work

Existing approaches to privacy protection primarily include methods targeting model memorization leakage and those addressing retrieval-phase privacy risks. To mitigate model memorization leakage, differential privacy (Li et al., 2022; Ponomareva et al., 2022) introduces carefully calibrated noise to training data, reducing the risk that sensitive information is memorized. Another approaches involve PII identification and anonymization techniques (Mouhammad et al., 2023; Subramani et al., 2023) preprocess training data by detecting and removing PII.

To address inference-phase privacy risks, synthetic data generation (Zeng et al., 2024b; Tang et al., 2024) replaces sensitive content with artificial data that maintains the statistical properties of real information. Data sanitization and anonymization (Dou et al., 2024b; Frikha et al., 2024; Staab et al., 2024) preprocess documents by removing or obfuscating sensitive information before they are fed into the model. Privacy-aware retrieval mechanisms (Huang et al., 2023; Arora et al., 2023) further modify the retrieval process to prevent privacy exposure while preserving enough public information. Details are provided in Appendix I

However, existing methods suffer from key limitations when applied to RAG. DP introduces noise that degrades model utility, making it unsuitable for RAG. PII anonymization, lacking dynamic privacy control, cannot adapt to scenario-specific definitions of privacy. Additionally, synthetic data generation requires modifying the retrieval corpus in advance, making database maintenance costly while potentially harming retrieval accuracy. Other methods like privacy-aware retrieval operate on isolated documents or queries, overlooking the risk of multi-document de-anonymization.

6 Conclusion

This paper introduces the Privacy Erasure Task and proposes Eraser4RAG, a novel model for controlled private knowledge erasure in RAG. We address key challenges including multi-document de-anonymization, controllable erasure, and maintaining downstream task utility. Our approach leverages triples to represent information and ensure fine-grained control over privacy erasure while minimizing unnecessary information loss. We further propose RL to effectively remove private triples while retaining essential public triples. In future work, we plan to avoid LLMs generating private information, potentially leveraging unlearning to differentiate private content during generation.

7 Limitations

Eraser4RAG still has limitations requiring further improvement. In some cases, relying on structured triples may not fully capture all information in text, limiting overall privacy preservation. To better capture sensitive information in unstructured text, future work should investigate more flexible approaches that do not rely solely on predefined triple structures. For example, end-to-end models that leverage contextual embeddings and dynamic attention mechanisms could be developed to identify and remove sensitive content directly from raw text.

References

Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. Reasoning over public and private data in retrieval-based systems. *Transactions of the Association for Computational Linguistics*, 11:902–921.

Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Sotirios K Goudos, Konstantinos E Psannis, Nikoleta Karditsioti, Theocharis Saoulidis, and Panagiotis Sarigiannidis. 2024. Evaluating the efficacy of ai techniques in textual anonymization: A comparative study. In *2024 7th International Balkan Conference on Communications and Networking (BalkanCom)*, pages 242–246. IEEE.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024a. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024b. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. In *Neurips Safe Generative AI Workshop 2024*.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv preprint arXiv:2407.02956*.

Anh-Tu Hoang, Barbara Carminati, and Elena Ferrari. 2023. Protecting privacy in knowledge graphs with personalized anonymization. *IEEE Transactions on Dependable and Secure Computing*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

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

Shouling Ji, Prateek Mittal, and Raheem Beyah. 2016. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, 19(2):1305–1326.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, and 1 others. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369.

Nina Mouhammad, Johannes Daxenberger, Benjamin Schiller, and Ivan Habernal. 2023. Crowdsourcing on sensitive data with privacy-preserving text rewriting. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 73–84, Toronto, Canada. Association for Computational Linguistics.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Retrieve, read and link: Fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. 2022. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv: Learning,arXiv: Learning*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Large language models are anonymizers. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024a. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524, Bangkok, Thailand. Association for Computational Linguistics.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. 2024b. Mitigating the privacy issues in retrieval-augmented generation (rag) via pure synthetic data. *arXiv preprint arXiv:2406.14773*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Experiment Details

PopQA retrieval documents are used for GPT-4o annotation and filtering, and the initial sampling rate for privacy triples is 25%. For SFT, we employ the AdamW optimizer (Loshchilov and Hutter, 2019) and train for 3 epochs with a learning rate of 5e-5. During RL phase, the discount factor $\gamma = 0.99$ and the learning rate is 1e-5. The hyperparameter p in the reward function R is initiated as 20 and increases by 5 per 350 iterations to 40. RL training data includes retrieval documents from PopQA, TriviaQA, NQ, and HotpotQA.

B Dataset Details

PopQA is a knowledge-intensive QA dataset with questions derived from structured knowledge bases. TriviaQA consists of trivia-style questions sourced from online trivia websites, featuring diverse topics. NQ contains real-world queries from Google Search. HotpotQA is a multi-hop QA dataset requiring reasoning over multiple documents. For the above dataset, use the provided training set to perform GPT-4o annotation and data filtering, and use the test dataset to conduct performance testing of the rewritten model. NewsQA a reading comprehension dataset built from CNN news articles.

C Entity Variations, Matching Heuristics, and Limitations

In real-world applications, entities may appear in various surface forms, including abbreviations, nicknames, or indirect references. **However, these various entities are precisely the challenge that researchers in knowledge extraction field aim to address, and our paper focuses on evaluating a model's ability to erase private knowledge**. To achieve this, we adapt the state-of-the-art model ReLiK (Orlando et al., 2024) from knowledge extraction field to ensure that the extraction tool can reliably and comprehensively identify knowledge from the documents. This model achieves state-of-the-art performance in both in-domain and out-of-domain benchmarks while using academic budget training and with up to 40x inference speed compared to competitors;

And we use the formal and well-structured Wikipedia corpus for our experiments. Entity references are typically consistent or only moderately abbreviated, e.g., "George James Rankin" may later appear as "George" or "George Rankin," but not as unrelated nicknames. This consistency allows relation extraction to work reliably, and we also implement a simple entity matching heuristic during triple classification: if all tokens of a short form appear in a longer one (i.e., every token in A is present in A'), we consider them as referring to the same entity. This helps prevent cases like <George James Rankin, works_at, TechVille> being labeled as private while <George, works_at, TechVille> remains in the public set, reducing the risk of inadvertently treating a private triple as unrelated to a semantically equivalent public one.

That said, we acknowledge that in more informal or noisy text domains, such as social media or medical records, surface forms of entities may vary significantly. In such scenarios, more sophisticated techniques like coreference resolution, alias detection, or embedding-based entity normalization would be required before relation extraction, to ensure robust privacy detection and erasure.

D Prompt Empirical Tuning

We experimented with several prompt variants before arriving at the current version, focusing on how well GPT-4o could minimize private triples while preserving the expression of public triples. In our tuning process, we found the following aspects to be particularly important for optimal performance:
(1) Step-by-step structure: We explicitly enumerated the rewriting objectives as numbered rules. This significantly improved GPT-4o's consistency and controllability.
(2) Explanation generation: Requiring the model to output an explanation of what public triples were preserved served as a form of self-verification. This helped us detect hallucinations or omissions in early drafts and encouraged more faithful preservation of public knowledge.
(3) Style anchoring: We emphasized that the rewriting should closely mirror the original style and phrasing of the source document when rewriting public content. Without this, GPT-4o tended to paraphrase unnecessarily, which could affect downstream utility.

To assess the quality of each prompt variant, we measured two key outcomes on a validation set: (1) the proportion of retained private triples in the rewritten text, and (2) the proportion of retained public triples in the rewritten text. The selected prompt yielded the best trade-off between these two, and its effectiveness is reflected in our quantitative results. We have added this point to the

revised manuscript.

E Split of Private and Public Triples

In the evaluations of Private Information Erasure, Public Information Preservation, and Resistance to De-Anonymization, the split of private and public triples follows the same random sampling procedure as in the training set. For the downstream RAG, we filter the private triples by removing any triple that explicitly represents a question–answer relationship, ensuring the queries are harmless. For the Resistance to Privacy Attacks evaluation, the private triples are directly constructed from the question–answer pairs.

F Implementation of Baselines

Abstraction. The original Abstraction approach is not designed to operate under a predefined private-public knowledge distinction. To make it applicable as a baseline in our privacy-focused setting, we adapted it by identifying the tail entities from each private triple as the target spans for anonymization. These entity spans were then fed into the anonymization model from Abstraction, which generates substitute spans intended to mask the identity of the original entities. Finally, we applied a regular-expression-based replacement procedure to substitute the entity mentions in the original text with the anonymized spans suggested by the model. This adaptation allowed us to evaluate Abstraction's anonymization capability under a controlled setting where privacy is explicitly defined, enabling a more direct comparison with our method.

Anonymizer. This work introduces an adversarial anonymization framework that leverages LLMs' strong inferential capabilities to detect and remove private information. Specifically, the method operates in two stages: (1) adversarial inference LLM first identifies potentially sensitive content in a text, and (2) anonymization LLM rewrites or removes such content to reduce the risk of disclosure while maintaining the overall utility of the text. In our experiments, we define privacy using the relation and tail entity to determine what information should be considered private. The anonymization process then aims to eliminate or obscure these private triples from the text, while preserving all other public knowledge.

Case 1:
Document A : Captain James Cook was a British explorer, navigator, cartographer, and captain in the Royal Navy .... James Cook joined the British merchant navy as a teenager and joined the Royal Navy in 1755. James Cook saw action in the Seven Years' War and subsequently surveyed and mapped much of the entrance... **Private**: [...<James Cook, participate, Seven Years ' War>...] **Public**: [<James Cook, country, British>, <James Cook, part of, Royal Navy>...]
Eraser4RAG: Captain James Cook was a British explorer, navigator, cartographer, and captain in the Royal Navy... James Cook joined the British merchant navy as a teenager and joined the Royal Navy in 1755. James Cook **saw action** and subsequently surveyed and mapped much of the entrance... **Abstraction**: Captain James Cook was a British explorer, navigator, cartographer, and captain in the Royal Navy... James Cook joined the British merchant navy as a teenager and joined the Royal Navy in 1755. James Cook **saw action in the conflicting event** and subsequently surveyed and mapped much of the entrance...
Document B : ...He saw action in the Seven Years' War, which was a significant conflict involving the British military. During this time, his contributions to the British forces were well recognized... **Private**: [...<Seven Years' War, country, British>...] **Public**: [... <James Cook, country, British>...]
Eraser4RAG: ... He saw action in **a significant conflict**. During this time, his contributions to the British forces were well recognized... **Abstraction**: ... He saw action in the **Seven Years' War**, which was a significant conflict involving a **famous country** military. During this time, his contributions to the country forces were well recognized...

Table 7: The case study comparing with Abstraction baseline.

Incognitext. Incognitext also employs two LLMs: one LLM infers the private attributes from a document, and the second LLM rewrites the document using incorrect private attribute values to mislead the adversarial LLM. The process involves two rounds of interaction to achieve anonymized rewriting of the document's private attributes. In our adaptation, for each document, we treat the relations in its private triples as the private attributes, the tail entities as the correct private attribute values, and use "unknown" as the incorrect private attribute values during rewriting.

NER. For each document, NER is performed using a BERT-based named entity recognition model to identify all entities in the text, which are then replaced with their corresponding identifiers.

G Case Study

In this case study shown in Table 7, we focus on the private relation: <James Cook, participate, Seven Years' War>, which is marked as private in Document A. In Document A, we successfully removes explicit mentions of James Cook's participation in the Seven Years' War in the rewritten Document A. The original phrase "James Cook saw action in the Seven Years' War" is omitted in the rewritten text, replaced with a more general statement such

Abstraction	gpt-each	gpt-all	Eraser4RAG
2.2633	1.7856	1.8659	0.8024

Table 8: Inference time of rewriting documents (s)

as "James Cook saw action and subsequently...," thus hiding the private relation. At the same time, public information like James Cook's nationality and affiliation with Royal Navy is preserved. In Document B, where <Seven Years' War, country, British> is private, the rewritten text changes the explicit mention "the Seven Years' War" to a vaguer "a significant conflict," aligning with the redacted portions of Document A, thereby concealing the private information while retaining the context of military involvement and contributions to British forces.

However, <James Cook, participate, Seven Years' War> is not explicitly defined as privacy in Doc B. For Abstraction, since privacy is considered only within individual documents rather than the global private triples, the result would include "Seven Years' War", without hiding <James Cook, participate, Seven Years' War>. Since Abstraction does not perform anonymization based on relationships, and British appears in both public and private relationships, Abstraction replaces British which should be treated as public information.

This case demonstrates that our method can effectively erase targeted private relations from documents while preserving as much public information as possible. It also highlights the importance of addressing privacy from a global, multi-document perspective, as private information might still be inferred if only single documents are considered independently.

H Inference Time

We evaluate the inference time required for different methods when rewriting documents. Since Eraser4RAG is based on the lightweight Flan-T5-large model, its inference time is significantly lower. This efficiency makes it more practical for real-world applications where computational cost is a concern. Moreover, the reduced inference time does not come at the expense of privacy preservation, as Eraser4RAG maintains strong privacy filtering capabilities while ensuring stable performance in downstream tasks.

system:
You are a relation assistant designed to rewrite any given text to remove information strictly related to the provided private triples. Each triple consists of a head entity, a relation, and a tail entity. Focus specifically on removing or generalizing relations between the head and tail entities in the provided private triples, without generalizing any other information arbitrarily.
Below, I will provide you with a text that needs to be anonymized, a set of public triples that contain non-private information, and a set of private triples that include private information. Your task is to rewrite the text to ensure that:
1. Analyze the text to identify any information that matches the private triples.
2. Only related tail entities in the private triples are replaced with the corresponding relations, while head entities should remain unchanged.
3. All relevant information from the public triples is preserved to maintain the accuracy and richness of the content.
4. The final text should be natural and coherent, closely mirroring the original text while excluding private relations and its tail entities.
5. Make sure to preserve any information that is not mentioned in the public and private triples.

user:
1. **Text**: [{text}]
2. **Public Triples**: {public_trps}
3. **Private Triples**: {private_trps}
After rewriting, provide the output in the following JSON format:
{ 'explanation': '<your explanation of the changes made>',
'anonymized_text': '<your rewritten text here>'}

Figure 5: The annotation prompt of GPT-4o.

I Related Work Details

Existing approaches to privacy protection primarily include methods targeting model memorization leakage and those addressing retrieval-phase privacy risks. To mitigate model memorization leakage, differential privacy (Li et al., 2022; Ponomareva et al., 2022) introduces carefully calibrated noise to training data, reducing the risk that sensitive information is memorized. For example, Li et al. (2022) fine-tunes models using task-aligned objectives, carefully tuned hyperparameters, and a memory-efficient gradient clipping technique called Ghost Clipping. Ponomareva et al. (2022) combines DP-SentencePiece, which adds noise to word counts for private tokenization, with DP-Training, which applies gradient clipping and noise during training. Another approaches involve PII identification and anonymization techniques (Mouhammad et al., 2023; Subramani et al., 2023) preprocess training data by detecting and removing personally identifiable information. Mouhammad et al. (2023); Subramani et al. (2023) use automated tools (Presidio (Mendels et al., 2018), Spacy (Honnibal and Montani, 2017)) and regular expression-based approaches to detect high-risk personal information in training corpora.

To address inference-phase privacy risks, synthetic data generation (Zeng et al., 2024b; Tang et al., 2024) replaces sensitive content with artificial data that maintains the statistical properties of real information. Zeng et al. (2024b) proposes a two-stage framework for generating synthetic data. The first stage uses attribute extraction and generation to preserve key information from the origi-

nal data throughout the database, while the second stage employs agent-based iterative refinement to enhance privacy. Tang et al. (2024) proposes to generate differentially private synthetic few-shot examples from private datasets. Data sanitization and anonymization (Dou et al., 2024b; Frikha et al., 2024; Staab et al., 2024) preprocess documents by removing or obfuscating sensitive information before they are fed into the model. Dou et al. (2024b) rephrases sensitive self-disclosure spans into less specific terms using fine-tuned language models to generate diverse alternatives. Staab et al. (2024) proposes a feedback-guided adversarial anonymization framework that leverages the inferential capabilities of LLMs to iteratively anonymize text. Privacy-aware retrieval mechanisms (Huang et al., 2023; Arora et al., 2023) further modify the retrieval process to prevent privacy exposure while preserving enough public information. Huang et al. (2023) decouples key and query encoders, isolating sensitive data exposure. Arora et al. (2023) restricts the use of private documents in subsequent public retrievals and ensures queries relying solely on private information remain local.

However, existing privacy protection methods suffer from key limitations when applied to RAG. DP introduces noise that degrades model utility, making it unsuitable for RAG, where precise knowledge retention is crucial. PII anonymization, lacking dynamic privacy control, cannot adapt to scenario-specific definitions of privacy. Additionally, synthetic data generation requires modifying the retrieval corpus in advance, making database maintenance costly while potentially harming retrieval accuracy due to distorted semantic relationships. Other methods like data sanitization and privacy-aware retrieval operate on isolated documents or queries, overlooking the risk of multi-document de-anonymization. Therefore, we propose Eraser4RAG to perform privacy erasure on retrieved documents. We utilize knowledge graph triples to represent information, which enables precise control over privacy removal while minimizing unnecessary knowledge loss. Our approach leverages the global privacy knowledge graph to effectively prevent multi-document de-anonymization risks.