

Dataset Protection via Watermarked Canaries in Retrieval-Augmented LLMs

Yepeng Liu^{1,†}, Xuandong Zhao², Dawn Song², Yuheng Bu¹

¹UC Santa Barbara ²UC Berkeley

[†]yepengliu@ucsb.edu

Abstract

Retrieval-Augmented Generation (RAG) has become an effective method for enhancing large language models (LLMs) with up-to-date knowledge. However, it may pose a risk of copyright infringement, as IP datasets may be incorporated into the knowledge database by malicious Retrieval-Augmented LLMs (RA-LLMs) without authorization. To protect the rights of the dataset owner, an effective dataset membership inference algorithm for RA-LLMs is needed. In this work, we introduce a novel approach, *CanaryTrace*, to safeguard the ownership of text datasets and effectively detect unauthorized use by RA-LLMs. Our approach preserves the original data completely unchanged while protecting it by inserting specifically designed canary documents into the IP dataset. These canary documents are created with synthetic content and embedded watermarks to ensure uniqueness, consistency, and statistical provability. During the detection process, unauthorized usage is identified by querying the canary documents and analyzing the responses of RA-LLMs for statistical evidence of the embedded watermark. Our experimental results demonstrate high query efficiency, detectability, and consistency, along with minimal perturbation to the original dataset, all without compromising the performance of the RAG system. Our code is available at <https://github.com/yepengliu/CanaryTrace>.

1 Introduction

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) with dynamic external knowledge, yet the potential usage of proprietary datasets raises security and intellectual property (IP) concerns. Unauthorized use of such data not only leads to IP infringement but also undermines economic efficiency (Ma, 2022). To address this, it is critical to develop protection strategies that identify unauthorized usage while preserving the dataset’s original utility. One effective approach

involves performing pre-release operations, such as inserting traceable markers, to safeguard the data owner’s rights in authorized RAG systems.

Specifically, embedding a watermark into the IP dataset and detecting it through the outputs of the trained LLMs offers an effective solution to address this issue. Wei et al. (2024) proposes inserting random character sequences into the IP dataset, enabling watermark detection by evaluating the loss of LLMs on these random character sequences. However, this approach embeds a visible watermark into the IP dataset, making it susceptible to being easily identified and removed (Liu et al., 2024c). Furthermore, the detection process relies on accessing the logits of LLMs, which are often inaccessible for proprietary models. To make the watermark more imperceptible, Jovanović et al. (2024) embeds an invisible watermark by paraphrasing each document in the IP dataset. However, the paraphrasing process may introduce significant distortions by altering the original dataset. Moreover, the capacity for watermark embedding varies across IP datasets, and *low-entropy* datasets often lack sufficient redundancy, making it hard to embed watermarks via paraphrasing without compromising their meaning or functionality.

These works highlight the importance of conducting imperceptible pre-release operations on IP datasets to enable effective detection. However, such operations must not compromise the original utility of the IP dataset in RAG. The most straightforward way to preserve utility is to keep the original data completely unchanged. Given that existing methods either compromise imperceptibility or significantly alter the original data, we ask: *Can we embed an imperceptible and effective watermark to reliably detect unauthorized usage while keeping the original data completely unchanged?*

In this paper, we formulate the problem as Dataset Membership Inference for Retrieval-Augmented Generation of LLMs (DMI-RAG) and

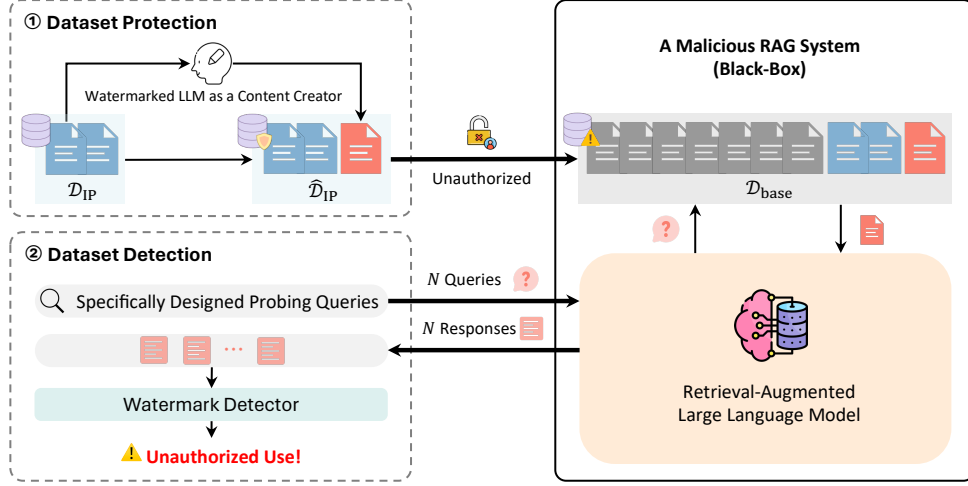


Figure 1: Overview of our *CanaryTrace*. In the dataset protection stage, we generate watermarked and synthetic canary documents based on the attributes of the documents from the IP dataset to form a protected dataset. A malicious RAG system may integrate an IP dataset into its base dataset without obtaining permission from the data owner. During the detection stage, the data owner can conduct black-box queries (without logits information) targeting these canary documents using specifically crafted questions and analyze the model responses to detect the presence of the watermark.

propose an effective solution. Our core idea is to embed a few watermarked *canaries*¹ into the IP dataset, enabling the dataset owner to verify their presence in suspected RA-LLMs. If the watermark embedded in the canaries is detected in the responses of RA-LLMs, the data owner can attribute their presence to unauthorized usage. This offers two advantages: 1) Preserving the original IP dataset maintains its quality and integrity, keeping its precise wording and style intact; 2) ensuring the reliable detection for low-entropy datasets: by carefully designing synthetic canaries, we can embed watermarks without altering the original dataset, ensuring effective detection.

Specifically, these canaries are designed to be inconspicuous to evade detection or removal by malicious RA-LLMs, while remaining distinctive enough to act as reliable markers for ownership verification. To achieve these properties, we use a watermarked LLM as a content creator to synthesize canary documents that align with the attributes of the IP dataset. This attribute-based synthesis ensures consistency between the canary documents and the IP dataset. The canary ensures uniqueness, minimizing semantic overlap with other documents in the RAG dataset and improving the accuracy of canary document retrieval. The watermarked LLM embeds an invisible watermark into the canary documents, allowing the watermark to diffuse into the

model responses when the canary documents are retrieved. Utilizing the robustness of the watermarking, the data owner can use the watermark in the responses as statistical evidence to detect unauthorized use of the protected dataset. In summary, our main contributions are listed as follows:

- We propose a dataset protection framework in a black-box setting that preserves the IP dataset’s documents intact while achieving high detection performance through LLM watermarking embedded in carefully designed canaries.
- We propose an automatic canary synthesis method that leverages the attributes of data from the IP dataset, ensuring high consistency with the original dataset.
- We conduct extensive experiments on various datasets and compare our method with baseline approaches. The results demonstrate that (1) our method achieves a 100% query accuracy for retrieving the canary documents on the NFCorpus dataset (Figure 4). (2) We achieve a 100% TPR@1%FPR with only 12 queries to the suspicious RAG system (Table 1). (3) Our method preserves the performance of downstream RAG tasks (Table 5). (4) We demonstrate that the proposed method maintains its effectiveness under scenarios where only a small subset of the protected dataset is stolen (Appendix B.1).

A comprehensive review of related work, including RAG, LLM watermarking techniques, data/

¹something acting as informers or decoys for the data owner.

dataset membership inference, and backdoor attacks, is provided in Appendix A.

2 Dataset IP Protection

2.1 Threat Model

Our threat model consists of two entities: the data owner, who seeks to protect datasets from unauthorized use, and a malicious RA-LLM, which owns a knowledge dataset and attempts to incorporate the IP dataset into its knowledge base without authorization. In our formulation, the data owner proactively performs operations on the IP dataset before its release, such as embedding invisible watermarks, to enable effective protection. The protected dataset is then accessible to authorized users or may be illegally obtained by unauthorized users. The data owner’s capability involves performing black-box dataset membership inference by inputting a limited number of crafted queries to the RA-LLM and analyzing only the generated responses. In the main paper, we assume that unauthorized users steal the entire dataset. However, in Appendix B.1, we also demonstrate that our method remains effective even when only a small subset of the IP dataset is stolen.

2.2 Problem Formulation

Our high-level idea is to embed watermarked canaries into the protected dataset, serving as evidence of unauthorized use and enabling reliable detection if the dataset is misappropriated. As shown in Figure 1, our method consists of two key components. In the dataset protection phase, the data owner generates watermarked canary documents and inserts them into the IP dataset before the release. In the dataset detection phase, the data owner performs multiple queries to the RA-LLM and provably identifies unauthorized use by framing it as a hypothesis testing problem to distinguish between the following two hypotheses:

- H_0 : The response does not contain watermark;
- H_1 : The response contains a watermark.

Canary Dataset Synthesis. Specifically, given a RA-LLM $\mathcal{M}_0$ with a knowledge dataset $\mathcal{D}_{\text{base}}$, and an IP dataset $\mathcal{D}_{\text{IP}}$. Our main idea is keeping original $\mathcal{D}_{\text{IP}}$ untouched, and insert a canary dataset $\mathcal{D}_s$ to get a protected dataset $\hat{\mathcal{D}}_{\text{IP}} = \mathcal{D}_{\text{IP}} \cup \mathcal{D}_s$, with the number of documents $|\mathcal{D}_s| \ll |\mathcal{D}_{\text{IP}}|$.

The inserted $\mathcal{D}_s$ should assist the data owner in enhancing query efficiency, providing statistical evidence, and remaining consistent with $\mathcal{D}_{\text{IP}}$.

Therefore, it should satisfy three key properties: uniqueness, consistency, and statistical provability. The uniqueness of the canary document minimizes semantic overlap with other documents in $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{IP}}$, thereby increasing the likelihood that it can be successfully retrieved using specifically designed queries. The consistency of the canary document ensures $\mathcal{D}_s$ shares similar attributes with the original IP dataset, making it more imperceptible. Statistical provability provides statistical evidence of the watermark across multiple queries, therefore enabling decisions with a controlled false alarm.

To achieve these properties, we design a canary data synthesis algorithm to automatically synthesize canary documents $\mathcal{S} : \mathcal{D} \mapsto \mathcal{D}_s \times \mathcal{Q}_s$, which uses a watermarked LLM to synthesize fictional text $d_s \in \mathcal{D}_s$ and the corresponding queries based on the attributes extracted from the sampled data $d \in \mathcal{D}_{\text{IP}}$. This ensures that the synthesized text retains uniqueness and relevance to the original dataset while embedding the necessary watermark for detection. Then, we synthesize query text to form the query dataset $q_s \in \mathcal{Q}_s$, specifically tailored for each $d_s \in \mathcal{D}_s$.

Dataset Membership Detection. A watermark detector, $D_w : \mathcal{Y}^* \mapsto \mathbb{R}$, is employed to analyze the outputs y of $\mathcal{M}_0$ generated in response to queries $q_s \in \mathcal{Q}_s$. Specifically, we can determine whether $\mathcal{D}_{\text{IP}} \subseteq \mathcal{D}_{\text{base}}$ if the detector satisfies the condition $D_w(y^{(1)} \oplus y^{(2)} \oplus \dots \oplus y^{(N)}) \geq \eta$, where η is a predefined threshold, N is the number of queries, $\oplus$ represents the concatenation operation.

2.3 Canary Dataset Synthesis

In this section, we present our canary dataset synthesis algorithm $\mathcal{S}$. As shown in Figure 2, it includes: 1) attributes extraction, 2) entity synthesis 3) description synthesis, 4) article synthesis, and 5) query question generation.

Attributes Extraction. To ensure consistency with $\mathcal{D}_{\text{IP}}$, each canary synthesis process begins by randomly sampling a document from $d \in \mathcal{D}_{\text{IP}}$. Then, we analyze and extract the key attributes $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ of sampled text using an attribute extraction function, attr_extr_func . This process is facilitated by an LLM $\mathcal{M}$ with appropriately designed prompts: $\mathcal{A} \leftarrow \mathcal{M}(\text{attr_extr_func}(d))$.

In our experiments, we extract four attributes from the sampled document d : $\{A_1 = \text{Topic}, A_2 = \text{Subtopic}, A_3 = \text{Writing Style}, A_4 = \text{Word Count}\}$. These attributes ensure that the syn-

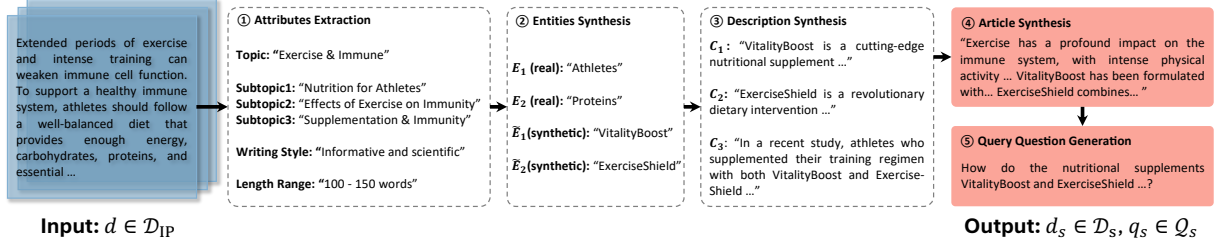


Figure 2: Workflow of our canary dataset synthesis algorithm. The process begins by randomly sampling a document from the IP dataset as a reference. Key attributes are then extracted from the reference document. Using these attributes, descriptions and relationships between synthetic entities are created. Finally, the algorithm outputs the synthetic text and a corresponding query question.

thesized documents align with the same topic, writing style, and length as the original dataset.

Entity Synthesis. The uniqueness of the canary documents significantly reduces overlap with other documents within $\mathcal{D}_{base}$ and $\mathcal{D}_{IP}$, ensuring a clear separation between the embeddings of the canary documents and those of other documents. This distinctiveness greatly enhances the accuracy of retrieving the target canary documents (Figure 4). To ensure the uniqueness of the synthesized text, we extract real entities $\mathcal{E} = \{E_1, \dots, E_n\}$ from d and create fictional entities $\tilde{\mathcal{E}} = \{\tilde{E}_1, \dots, \tilde{E}_n\}$ conditioned on the attributes $\mathcal{A}$ using the LLM with the `ent_synth_func`: $\{\mathcal{E}, \tilde{\mathcal{E}}\} \leftarrow \mathcal{M}(\text{ent_synth_func}(d, \mathcal{A}))$.

Description Synthesis. Then, we synthesize the fictional descriptions for each entity in $\tilde{\mathcal{E}}$ and the relations between them $\mathcal{C} = \{C_1, \dots, C_n\}$ using the LLM with the `des_synth_func`: $\mathcal{C} \leftarrow \mathcal{M}(\text{des_synth_func}(\tilde{\mathcal{E}}, \mathcal{A}))$. This step preserves contextual relevance to $\mathcal{D}_{IP}$ while ensuring that synthesized text remains distinct and identifiable.

Article Synthesis. Finally, we collect the synthesized descriptions $\mathcal{C}$ to generate a new article d_s using the LLM with `article_synth_func`:

$$d_s \leftarrow \mathcal{M}_w(\text{article_synth_func}(\mathcal{C}, \mathcal{A})), \quad (1)$$

where $\mathcal{M}_w$ is a watermarked LLM, which will be detailed in Section 2.4.

Query Question Generation. For each $d_s \in \mathcal{D}_s$, we prompt an LLM to generate a question that can only be answered by reading d_s using `query_synth_func`:

$$q_s \leftarrow \mathcal{M}(\text{query_synth_func}(d_s)) \quad (2)$$

The generated q_s is used to retrieve the corresponding d_s from the suspicious RA-LLM during the detection process. All the specific prompts used

in our experiments are provided in Appendix D. Examples of the canary documents are provided in Appendix E.

While synthesizing fictional canary documents can improve retrieval accuracy, it also raises the risk of being identified and removed by malicious RA-LLMs that potentially use fact-checking tools. To alleviate this risk, we introduce a canary pre-filtering method. Specifically, we use a recently released LLM as a powerful fact-checker to screen out more suspicious canaries, ensuring that only those resistant to fact-checking are retained. More details are discussed in Appendix B.2.

2.4 Watermarking Synthetic Dataset

The process outlined in Section 2.3 introduces the uniqueness and consistency of the synthesized dataset. In this section, we detail the approach for ensuring the provability by integrating a watermarked LLM $\mathcal{M}_w$ into (1) of the algorithm $\mathcal{S}$.

We employ the green-red list watermarking scheme (Zhao et al., 2023), which randomly partitions the LLM’s vocabulary $\mathcal{V}$ into the global fixed green (G) and red (R) lists, such that $|G| = \gamma|\mathcal{V}|$ and $|R| = (1 - \gamma)|\mathcal{V}|$, where $\gamma \in (0, 1)$ represents the proportion of tokens assigned to the green list. This scheme slightly increases the logits ℓ of tokens in the green list at each time step, such that $\hat{\ell}_t[k] \leftarrow \ell_t[k] + \delta$ for $k \in G$. Therefore, the output of the watermarked LLM will exhibit a higher probability of green tokens. The key assumption here is that the watermark embedded in the document should remain in the output (Sander et al., 2024) of the RA-LLM (which is shown in Section 3) when the watermarked document is retrieved in response to a query. Therefore, it provides a way to determine the dataset membership by analyzing the model’s response and calculating the probability of green tokens present in the output.

2.5 Dataset Membership Inference

By applying the canary dataset synthesis algorithm $\mathcal{S}$, we can obtain the query dataset $\mathcal{Q}_s$ and canary dataset $\mathcal{D}_s$. The data owner can combine $\mathcal{D}_{IP}$ and $\mathcal{D}_s$ to construct the new dataset $\hat{\mathcal{D}}_{IP}$ prior to its release. During the dataset membership inference process, for a suspicious RA-LLM $\mathcal{M}_0$, we first query the model N times using different $q_s^{(i)} \in \mathcal{Q}_s$ and obtain the response $y^{(i)} = \mathcal{M}_0(\mathcal{R}(q_s^{(i)}, \mathcal{D}_{base}))$. Based on the discussion in Section 2.4, the dataset membership inference problem can be effectively transformed into a watermark detection problem. However, directly detecting the watermark in a single response is challenging because the embedded watermark in d_s becomes significantly weakened after passing through the RAG pipeline. Therefore, we concatenate N responses into a single sequence, denoted as $\mathbf{y} = y^{(1)} \oplus y^{(2)} \oplus \dots \oplus y^{(N)}$, to enhance the statistical detectability of the watermark in responses.

Specifically, we detect the watermark by computing the z-statistic of the response $\mathbf{y}$, i.e.,

$$z_{\mathbf{y}} = (|\mathbf{y}|_G - \gamma T) / \sqrt{\gamma(1 - \gamma)T}, \quad (3)$$

where T is the number of tokens in $\mathbf{y}$, and $|\mathbf{y}|_G$ is the count of the tokens in green list. The response $\mathbf{y}$ is identified as watermarked if $z_{\mathbf{y}} > \eta$, where η is a pre-defined threshold. Therefore, we can equivalently conclude if $\mathcal{D}_{IP} \subseteq \mathcal{D}$. Moreover, we establish a theoretical guarantee that controls the false-alarm rate, thereby limiting the risk of falsely accusing benign RA-LLMs. Details can be found in Appendix C.

3 Experiments

3.1 Experiment Setting

Implementation Details. (1) *Model Settings:* For our canary dataset synthesis algorithm, we use GPT-4o-mini to extract attributes and generate fictional descriptions. Additionally, we use watermarked Llama-3.1-70B-Instruct to embed the watermark into the synthesized articles. (2) *Watermark Settings:* For the watermarking algorithm, we set $\gamma = 0.5$ and $\delta = 2.0$. (3) *Retriever Settings:* For retriever, we conduct experiments based on the Contriever-ms (Izacard et al., 2021) and ANCE (Xiong et al., 2020). We select the top $K = 3$ most relevant documents based on the cosine similarity. In our main experiments, each canary document is queried only once with a

single question. We investigate the detection performance under the setting where multiple queries are conducted per canary document in Appendix B.4. All experiments are conducted on NVIDIA A100 GPU.

Baselines. We compare our methods with two existing dataset membership inference methods leveraging watermark, including Ward (Jovanović et al., 2024), and WWJ (Wei et al., 2024). For Ward, consistent with its default settings, we use KGW (Kirchenbauer et al., 2023) to paraphrase the whole dataset and set $\gamma = 0.5$, $\delta = 3.5$, and window size $h = 2$. We adapt WWJ to our setting; implementation details can be found in Appendix B.3.

Datasets. We use the MS MARCO (Nguyen et al., 2016) as the knowledge dataset of RA-LLMs, which is a large-scale real-world web document corpus. In the main results, we set the size of the base dataset to 200,000. To further demonstrate the effectiveness of our method, we extend this size to 2 million in Appendix B.5. In this setting, we assume that potential semantic overlap may exist between the knowledge dataset and the IP dataset, which often happens in the real world. We evaluate our performance using the biomedical information retrieval dataset, NFCorpus (Boteva et al., 2016), as the IP dataset. Moreover, we use CQADupStack-Mathematica (Hoogeveen et al., 2015) as IP dataset to evaluate the detection performance of our methods on the low-entropy dataset. We evaluate the downstream performance of different methods on Chinese-poem² and DROP dataset (Dua et al., 2019).

Evaluation Metrics. To evaluate and compare the performance of different methods, we employ the following evaluation metrics.

- We use Target Retrieval Accuracy to evaluate the proportion of corresponding watermarked documents successfully retrieved by the queries. It is computed as: $\frac{1}{N} \sum_{i=1}^N \mathbb{1}(d^{(i)} \in \mathcal{R}(q^{(i)}))$.
- For detectability, the detection performance is assessed using the ROC-AUC value, which measures the ability of a detector to distinguish between classes by evaluating the trade-off between true positive rate (TPR) and false positive rate (FPR) across different thresholds. Moreover, we report detection performance at different FPR values, such as TPR@1%FPR and TPR@10%FPR.
- We evaluate the dataset distortion caused by different dataset protection methods using BLEU

²https://huggingface.co/datasets/xmj2002/tang_poems

Table 1: Detection performance across different methods on NFCorpus dataset with varying query quota. Our method achieves 100% detection performance with only 12 query quota.

Query Quota	WWJ			Ward			Ours		
	ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR
2	-	-	-	0.809	0.274	0.482	0.970	0.546	0.908
4	0.986	0.968	0.972	0.880	0.336	0.660	0.995	0.850	0.996
6	0.987	0.969	0.974	0.922	0.530	0.762	0.998	0.960	1.000
8	0.990	0.970	0.974	0.940	0.520	0.820	0.999	0.994	1.000
10	0.990	0.972	0.976	0.963	0.660	0.900	0.999	0.998	1.000
12	0.991	0.972	0.978	0.982	0.792	0.954	1.000	1.000	1.000

(Papineni et al., 2002) and MAUVE (Pillutla et al., 2021) scores.

- The consistency is measured using Perplexity and QuRating (Wettig et al., 2024), which are two methods to curate a dataset. More details can be found in Appendix B.3.
- We calculate the response correctness using GPT-4o-mini as a judge to assess the impact of each dataset protection method on the downstream performance of RA-LLMs.

3.2 Main Results

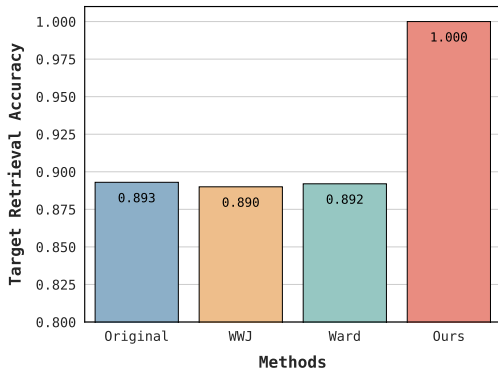


Figure 4: Retrieval Accuracy for different methods. Our method achieves 100% Retrieval Accuracy.

Target Retrieval Accuracy. Target Retrieval Accuracy is a crucial performance metric for the DMI-RAG task. A high Target Retrieval Accuracy ensures that the target watermarked documents can be retrieved with high probability using specifically designed queries. In this experiment, Target Retrieval Accuracy is calculated using 500 different questions to retrieve the corresponding watermarked documents for each method. Specifically, for our method, we insert 500 canary documents into the IP dataset. Figure 4 shows the Target Retrieval Accuracy for the original IP dataset and different baselines. The ‘Original’ represents the original IP dataset without any modification, where we generate 500 different questions to query different documents in the original IP dataset. As

shown in the table, our method outperforms all others, achieving a perfect Retrieval Accuracy of 100%. This is attributed to the insertion of synthetic fictional canaries, which minimizes the potential semantic repetition with other documents in the base dataset. In contrast, the methods employed by WWJ and Ward affect the Target Retrieval Accuracy to varying degrees compared to the original IP dataset.

Detectability. We evaluate and compare the detection performance across different query quotas for various methods. Our goal is to achieve superior detection performance with fewer number of queries. As shown in Table 1, our method outperforms the baseline approaches by achieving superior detection performance with a lower query count. In particular, we achieve 100% TPR@1%FPR by only querying the RA-LLM 12 times, while Ward is only 79.2% on the same quota. One main reason is that our method achieves a high Target Retrieval Accuracy.

Table 2: Dataset distortion for different methods. Our method achieves the lowest dataset distortion for both BLEU and MAUVE.

Methods	WWJ	Ward	Ours
BLEU (↑)	0.981	0.132	0.997
MAUVE (↑)	0.004	0.340	0.999

Dataset Distortion. Dataset distortion quantifies the extent of changes introduced by various methods to the original dataset. To evaluate this, we use BLEU and MAUVE scores. BLEU assesses the n -gram word-level overlap between the original dataset and the protected dataset, while MAUVE evaluates the similarity in distributions between the two datasets. As shown in Table 2, our method achieves the highest BLEU and MAUVE scores compared to other methods, demonstrating minimal distortion to the original dataset.

This is attributed to our approach of keeping the original documents entirely unchanged while en-

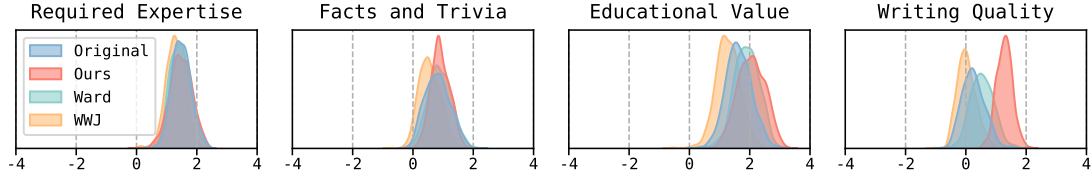


Figure 3: Distribution of quality ratings using QuRating across four aspects for various methods.

suring that the inserted canary documents share the same attributes as the original IP dataset. However, Ward modifies the expressions and words in the original dataset, resulting in low BLEU and MAUVE scores. The random sequence inserted by WWJ does not alter the original expressions, resulting in a high BLEU score. However, it introduces a significant distribution shift between the original IP dataset and the protected dataset, leading to a low MAUVE score.

Consistency. The consistency of our method is evaluated using two dataset curation methods: perplexity and QuRating. For perplexity, we calculate the perplexity of the canary documents generated by our method, the paraphrased documents produced by Ward, and the documents with inserted random sequences created by WWJ. We set the maximum perplexity value of the original dataset as a threshold to compute the Filtering Rate. In particular, we compute the perplexity using GPT-3. Table 3 shows that our canary documents exhibit a low average perplexity, comparable to that of the original IP dataset, and a 0 Filtering Rate. This demonstrates that our synthetic data is difficult to detect and remove based on the perplexity. However, WWJ shows a high Filtering Rate and average perplexity. This is because the randomly sampled sequences are inconsistent with the natural language patterns expected by a language model, resulting in extremely high perplexity.

For QuRating, we measure the canary documents generated by our method, the paraphrased documents produced by Ward, and the documents with inserted random sequences created by WWJ from four different aspects, obtaining the corresponding scores for each. The dataset distributions across different aspects are shown in Figure 3. As can be seen, our canary documents exhibit higher writing quality, making them less likely to be removed based on this criterion. For the remaining three aspects, our synthetic documents exhibit a similar distribution to the original documents, demonstrating their consistency with the original IP dataset. Both perplexity and QuRating demon-

strate the high quality and consistency of our canary documents.

Table 3: Perplexity and Filtering Rate across different methods. Our method achieves a perplexity comparable to that of the original dataset and 0% Filtering Rate.

Methods	Original	WWJ	Ward	Ours
Avg Perplexity ($\downarrow$)	6.736	12.734	9.453	6.952
Filtering Rate ($\downarrow$)	-	0.104	0.015	0.000

Downstream Performance. We further investigate the impact of different methods on the downstream performance by measuring the correctness of the model response. We use two datasets for two different tasks.

For DROP, we evaluate the correctness of the model’s responses on the protected dataset, specifically for discrete reasoning and numerical computation tasks. As shown in Table 5, our method maintains the correctness of the original dataset. This is because our method preserves the original data without any modifications. In contrast, the correctness of Ward decreases from 0.772 to 0.767. The drop in Ward’s performance suggests that paraphrasing may introduce additional errors, negatively affecting downstream tasks.

Table 5: Downstream Performance on two datasets.

Methods	Dataset	Correctness
Original	DROP	0.772
	Chinese-Poem	0.908
WWJ	DROP	0.771
	Chinese-Poem	0.900
Ward	DROP	0.767
	Chinese-Poem	0.719
Ours	DROP	0.772
	Chinese-Poem	0.908

Moreover, for literary datasets, such as poetry, we evaluate correctness in terms of poem appreciation. Specifically, we emphasize the critical importance of preserving the original data intact, as even slight modifications to the words in a poem can significantly alter the author’s intended expression and emotional tone, thereby impacting the overall inter-

Table 4: Detection performance for different methods on low-entropy dataset (Mathematica) with varying query quota. Our method achieves 100% detection performance with a query quota of just 10, while Ward shows a significantly lower detection performance.

Query Quota	Ward			Ours		
	ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR
2	0.572	0.029	0.116	0.975	0.584	0.938
4	0.614	0.016	0.215	0.997	0.920	1.000
6	0.623	0.037	0.244	0.999	0.974	1.000
8	0.628	0.012	0.228	0.999	0.998	1.000
10	0.668	0.049	0.219	1.000	1.000	1.000
12	0.680	0.041	0.235	1.000	1.000	1.000

pretation and appreciation of the work. From Table 5, the correctness of Chinese poem appreciation for Ward drops significantly from 0.908 to 0.719, whereas our method maintains the same level of correctness.

3.3 Detection Performance on Hard Conditions

Detectability on Low Entropy IP Dataset. We further evaluate the detection performance of our method on a low-entropy IP dataset, CQADupStack-Mathematica (Hoogeveen et al., 2015), and compare it with Ward. As shown in Table 4, our method maintains a high detection performance, achieving 100% TPR@1%FPR using only 10 query quotas. In contrast, Ward achieves significantly lower detection performance, with 0.049 TPR@1%FPR at the same query quota. This is because low-entropy IP datasets lack redundancy, making it difficult to perform modifications. As a result, paraphrasing can embed only a limited amount of watermark, which fails to persist in the model’s responses. In contrast, our method remains unaffected, as we embed watermarks into carefully designed canaries, which are not constrained by the entropy of the original IP dataset.

Table 6: Detection performance on hard system prompt.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
2	0.962	0.408	0.894
6	0.997	0.868	0.996
10	0.999	0.996	1.000
14	1.000	1.000	1.000

Hard System Prompt. We investigate the detection performance under two types of system prompts: easy and hard. An easy system prompt is straightforward, asking the model to answer the question without imposing additional constraints or complex instructions. A hard system prompt includes more restrictive instructions, such as strictly avoiding verbatim text or excessive paraphrasing.

Table 1 shows detection performance under the easy system prompt. In Table 6, we showcase the detection performance under the hard system prompt. The results demonstrate that our method achieves the 100% TPR@1%FPR with 14 query quotas, indicating that the detection performance remains unaffected even with hard system prompt.

Additional Main Results. In Appendix B.4, we present two **extra results**: (1) Detection performance when querying a single canary document multiple times. Specifically, in the multiple queries per canary document setting, we show that the data owner can achieve 100% detection performance with as few as 5 canary documents. (2) Robustness of our method under the canary paraphrasing attack and watermarked RA-LLM.

Ablation Study. In Appendix B.5, we conduct the following **ablation studies**: (1) The impact of watermark strength on detection performance. (2) Detection performance on different RAG systems. (3) Detection performance with a larger base dataset. (4) Detection performance when $K = 1$. (5) In Appendix B.6, we use a different watermarking approach (Kirchenbauer et al., 2023) to generate the watermarked canaries, further demonstrating the effectiveness of our approach. (6) Since our canary synthesis algorithm is capable of generating highly realistic canaries, we also examine this scenario and analyze its impact on retrieval accuracy.

4 Conclusion

In this paper, we propose a novel method to protect proprietary datasets from unauthorized use while preserving their original integrity by embedding carefully designed canary documents into the IP-protected dataset. Our experimental results demonstrate the effectiveness of our method in both detectability and consistency, making it a practical and reliable approach for real-world applications.

Limitations

While our proposed method demonstrates strong performance in protecting the dataset’s IP with minimal distortion, our method primarily focuses on text-based RAG systems. Although text is currently the most prevalent modality for RAG applications, the rapid development of multimodal models suggests a growing need for multimodal IP protection. Extending our canary synthesis and watermarking pipeline to multimodal datasets (e.g., synthesizing and embedding watermarks into image or audio canaries) remains an open area for future exploration.

References

- Li An, Yujian Liu, Yepeng Liu, Yuheng Bu, Yang Zhang, and Shiyu Chang. 2026. A reinforcement learning framework for robust and secure llm watermarking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198.
- Li An, Yujian Liu, Yepeng Liu, Yang Zhang, Yuheng Bu, and Shiyu Chang. 2025. Defending llm watermarking against spoofing attacks with contrastive representation learning. *arXiv preprint arXiv:2504.06575*.
- Maya Anderson, Guy Amit, and Abigail Goldsteen. 2024. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*.
- Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.
- Souradip Chakraborty, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2024. Position: On the possibilities of ai-generated text detection. In *Forty-first International Conference on Machine Learning*.
- Hongyan Chang, Hamed Hassani, and Reza Shokri. 2024a. Watermark smoothing attacks against language models. *arXiv preprint arXiv:2407.14206*.
- Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. 2024b. Context-aware membership inference attacks against pre-trained large language models. *arXiv preprint arXiv:2409.13745*.
- Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.
- Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. 2025. Improved unbiased watermark for large language models. *arXiv preprint arXiv:2502.11268*.
- Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *arXiv preprint arXiv:2407.12784*.
- Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. 2024. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*.
- Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.
- Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. 2024. Gumbelsoft: Diversified language model watermarking via the gumbelmax-trick. *arXiv preprint arXiv:2402.12948*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Eva Giboulot and Teddy Furon. 2024. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *arXiv preprint arXiv:2403.04808*.
- Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. 2024. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12374–12384.
- Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Heng Huang, and 1 others. 2024. Zeromark: Towards dataset ownership verification without disclosing watermark. *Advances in Neural Information Processing Systems*, 37:120468–120500.

- Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Chenxi Liu, Yanshuo Chen, and Heng Huang. 2025. Towards copyright protection for knowledge bases of retrieval-augmented language models via reasoning. *arXiv preprint arXiv:2502.10440*.
- Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. 2023. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *Advances in Neural Information Processing Systems*, 36:54421–54450.
- Haiyun He, Yepeng Liu, Zhuoer Shen, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2026. Fundamental trade-offs in multi-bit watermarking of stochastic processes. *arXiv preprint arXiv:2605.08826*.
- Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2024. Universally optimal watermarking schemes for llms: from theory to practice. *arXiv preprint arXiv:2410.02890*.
- Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. 2025. Distributional information embedding: A framework for multi-bit watermarking. *arXiv preprint arXiv:2501.16558*.
- Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.
- Jiahao Huo, Shuliang Liu, Bin Wang, Junyan Zhang, Yibo Yan, Aiwei Liu, Xuming Hu, and Mingxun Zhou. 2025. Pmark: Towards robust and distortion-free semantic-level watermarking with channel constraints. *arXiv preprint arXiv:2509.21057*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. 2024. Ward: Provable rag dataset inference via llm watermarks. *arXiv preprint arXiv:2410.03537*.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.
- Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. 2024a. Robust detection of watermarks for large language models under human edits. *arXiv preprint arXiv:2411.13868*.
- Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. 2024b. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *arXiv preprint arXiv:2404.01245*.
- Yiming Li, Shuo Shao, Yu He, Junfeng Guo, Tianwei Zhang, Zhan Qin, Pin-Yu Chen, Michael Backes, Philip Torr, Dacheng Tao, and 1 others. 2025. Rethinking data protection in the (generative) artificial intelligence era. *arXiv preprint arXiv:2507.03034*.
- Yuying Li, Gaoyang Liu, Yang Yang, and Chen Wang. 2024c. Seeing is believing: Black-box membership inference attacks against retrieval augmented generation. *arXiv e-prints*, pages arXiv–2406.
- Aiwei Liu, Leyi Pan, Xuming Hu, Shu’ang Li, Lijie Wen, Irwin King, and Philip S Yu. 2023a. An unforgeable publicly verifiable watermark for large language models. *arXiv preprint arXiv:2307.16230*.
- Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models, 2024. <https://arxiv.org/abs/2310.06356>.
- Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024a. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36.
- Mingrui Liu, Sixiao Zhang, and Cheng Long. 2024b. Mask-based membership inference attacks for retrieval-augmented generation. *arXiv preprint arXiv:2410.20142*.
- Xiaoqun Liu, Jiacheng Liang, Muchao Ye, and Zhao-han Xi. 2024c. Robustifying safety-aligned large language models through clean data curation. *arXiv preprint arXiv:2405.19358*.
- Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*.
- Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. 2025. Image watermarks are removable using controllable regeneration from clean noise. In *International Conference on Learning Representations*, volume 2025, pages 87310–87327.

- Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. 2026a. In-context watermarks for large language models. In *International Conference on Learning Representations*, volume 2026, pages 53136–53162.
- Yepeng Liu, Xuandong Zhao, Dawn Song, Gregory W Wornell, and Yuheng Bu. 2026b. Position: Llm watermarking should align stakeholders’ incentives for practical adoption. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25887–25903.
- Yixin Liu, Hongsheng Hu, Xuyun Zhang, and Lichao Sun. 2023b. Watermarking text data on large language models for dataset copyright protection. *arXiv preprint arXiv:2305.13257*.
- Qian Lou, Yepeng Liu, and Bo Feng. 2023. Trojtext: Test-time invisible textual trojan insertion. *arXiv preprint arXiv:2303.02242*.
- Peizhuo Lv, Mengjie Sun, Hao Wang, Xiaofeng Wang, Shengzhi Zhang, Yuxuan Chen, Kai Chen, and Limin Sun. 2025. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 1709–1723.
- Yueyuan Ma. 2022. Specialization in a knowledge economy. Available at SSRN, 4052990.
- Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092.
- Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.
- Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, and 1 others. 2024a. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*.
- Leyi Pan, Aiwei Liu, Yijian Lu, Zitian Gao, Yichen Di, Shiyu Huang, Lijie Wen, Irwin King, and Philip S Yu. 2024b. Waterseeker: Pioneering efficient detection of watermarked segments in large documents. *arXiv preprint arXiv:2409.05112*.
- Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. 2025. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25002–25009.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.
- Zhenting Qi, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. 2024. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. *arXiv preprint arXiv:2402.17840*.
- Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. 2024. Watermarking makes language models radioactive. *arXiv preprint arXiv:2402.14904*.
- Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. 2024. Proving membership in llm pre-training data via data watermarks. *arXiv preprint arXiv:2402.10892*.
- Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. QuRating: Selecting high-quality data for training language models. In *International Conference on Machine Learning (ICML)*.
- Yihan Wu, Ruibo Chen, Zhengmian Hu, Yanshuo Chen, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. Distortion-free watermarks are not truly distortion-free under watermark key collisions. *arXiv preprint arXiv:2406.02603*.
- Yihan Wu, Ruibo Chen, Georgios Milis, and Heng Huang. 2025a. An ensemble framework for unbiased language model watermarking. *arXiv preprint arXiv:2509.24043*.
- Yihan Wu, Xuehao Cui, Ruibo Chen, and Heng Huang. 2025b. Analyzing and evaluating unbiased language model watermark. *arXiv preprint arXiv:2509.24048*.
- Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. 2023. Dipmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.
- Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2023. Trojllm: A black-box trojan prompt attack on large

- language models. *Advances in Neural Information Processing Systems*, 36:65665–65677.
- Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. 2024. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*.
- Jingqi Zhang, Ruibo Chen, Yingqing Yang, Peihua Mai, Heng Huang, and Yan Pang. 2025a. Leave no trace: Black-box detection of copyrighted dataset usage in large language models via watermarking. *arXiv preprint arXiv:2510.02962*.
- Yu Zhang, Shuliang Liu, Xu Yang, and Xuming Hu. 2025b. Catmark: A context-aware thresholding framework for robust cross-task watermarking in large language models. *arXiv preprint arXiv:2510.02342*.
- Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.
- Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, and 1 others. 2024a. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*.
- Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2024b. Permute-and-flip: An optimally robust and watermarkable decoder for llms. *arXiv preprint arXiv:2402.05864*.
- Chaoyi Zhu, Jeroen Galjaard, Pin-Yu Chen, and Lydia Y Chen. 2024. Duwak: Dual watermarks in large language models. *arXiv preprint arXiv:2403.13000*.
- Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

A Background and Related Work

Retrieval Augmented Generation. RAG (Karpukhin et al., 2020; Xiong et al., 2020; Lewis et al., 2020) integrates information retrieval with natural language generation to enhance the response quality and relevance (Gao et al., 2023). In general, a RAG system involves three key components: the retriever $\mathcal{R}$, the generator $\mathcal{M}_0$, and a knowledge database $\mathcal{D}_{\text{base}}$. Given a query q , the retriever maps the input and documents in the knowledge database into embeddings within the same space. It then searches the knowledge database to retrieve K most relevant documents, $\{d_1, \dots, d_K\} = \mathcal{R}(q, \mathcal{D}_{\text{base}})$, based on the distance metric like cosine-similarity. In the generation phase, the generator produces the response given the query and the retrieved documents, i.e., $y = \mathcal{M}_0(d_1, \dots, d_K, q)$.

LLMs Watermarking. LLM watermarking embeds watermark into the text throughout the entire generation process (Li et al., 2024b; He et al., 2024; Li et al., 2024a; Zhao et al., 2024a,b; Fu et al., 2024; Giboulot and Furon, 2024; Wu et al., 2023, 2025a,b; He et al., 2025, 2026; Wu et al., 2024; Chen et al., 2025; Chakraborty et al., 2024; Chang et al., 2024a; Liu et al., 2024a; Pan et al., 2024a,b; Liu et al., 2023a; Zhu et al., 2024; Huo et al., 2025; Zhang et al., 2025b; Liu et al., 2025, 2026a,b; An et al., 2025, 2026). This is typically achieved by either perturbing the logits of the LLM (Liu and Bu, 2024; Kirchenbauer et al., 2023; Zhao et al., 2023) or manipulating the sampling process (Kuditipudi et al., 2023; Christ et al., 2024). In particular, robustness is a critical property of LLM watermarking, enabling the detection of watermarks even after significant text modifications. For the DMI-RAG task, a robust watermark is essential to ensure its persistence from the watermarked document to the response of the RAG system. Therefore, among all existing LLM watermarking methods, we adopt the watermarking by (Zhao et al., 2023) due to its simplicity and provable robustness.

Data Membership Inference. Data membership inference (Chang et al., 2024b; Qi et al., 2024) for RAG (Liu et al., 2024b; Lv et al., 2025) aims to determine whether a specific instance is included in the knowledge dataset of RA-LLMs. Li et al. (2024c) introduces a gray-box approach that computes a membership score by combining text-data similarity with generation perplexity. Anderson et al. (2024) propose a method that uses a spe-

cially designed prompt to black-box ask whether a dataset member appears in the context and deduces the membership status based on the model’s answer.

Backdoor Attack. For dataset membership inference in RAG, the dataset owner can proactively perform specific operations on the original dataset to gather evidence of unauthorized usage. Backdoor attack (Chaudhari et al., 2024; Cheng et al., 2024; Chen et al., 2024; Lou et al., 2023; Xue et al., 2023) is an approach used to identify unauthorized usage by embedding triggers in the dataset. These triggers cause RA-LLMs to produce specified abnormal responses when queried with certain inputs. Zou et al. (2024) proposes injecting malicious texts into the dataset to manipulate RA-LLMs into generating a predetermined incorrect answer for a specific question. In certain respects, our method shares similarities with a backdoor. However, unlike existing backdoor attacks, which compromise the functionality or performance of the original dataset, our approach is designed to avoid such issues.

Dataset Membership Inference. Dataset Membership Inference (Maini et al., 2024; Zhang et al., 2024; Maini et al., 2021; Guo et al., 2023, 2025, 2024; Zhang et al., 2025a; Li et al., 2025; Golatkar et al., 2024) for protection can be achieved through proactive watermark embedding (Liu et al., 2023b; Panaitescu-Liess et al., 2025). Wei et al. (2024) proposes inserting a random sequence repeatedly into the dataset and then computing the loss of the suspicious LLM on this sampled sequence to determine the dataset’s presence. Most recently, Jovanović et al. (2024) proposes to use a watermarked LLM to paraphrase each document in the IP-protected dataset and detect the watermark in the responses of RA-LLM. In this paper, we embed *invisible* watermarks into a small number of carefully designed canaries and insert them into the IP-protected dataset. Our approach preserves the original dataset untouched while achieving high detectability.

B Additional Experiment Results

B.1 Discussion about only a subset of IP dataset is stolen

Although we assume the entire IP dataset is stolen in the threat model, our method remains effective even when only a subset of the IP dataset is stolen, as long as a sufficient number of canaries are in-

sorted. Considering a realistic setting where we could query one document multiple times (Appendix B.4), we could realize 100% TPR@1%FPR with 5 canaries. Assuming $|D_{IP}| = 100,000$, the stolen dataset includes at least 5 canaries with 99% probability. We compute the minimum canary set size $|D_s|$ needed for various stolen proportions. As shown in the Table 7, inserting just 0.054% canaries ensures, with 99% probability, that at least 5 will be captured, even if only 20% of the dataset is leaked.

Table 7: Minimum canary set percentage needed for various stolen proportions.

	20%	40%	60%	80%
$ D_s / D_{IP} $	5.4×10^{-4}	2.5×10^{-4}	1.5×10^{-4}	1.0×10^{-4}

B.2 Canary pre-filtering Setting

To filter out highly suspicious canaries and retain only those resistant to fact-checking, we employed a recently released LLM, Gemini-2.5-Flash-Lite (Comanici et al., 2025), as an automated and powerful fact-checker. It simulates the behavior of malicious RA-LLMs that may leverage fact-checking tools to conduct attacks. Specifically, we synthesized 500 distinct canary documents and asked the model to flag any text containing fictional entities. Of these, 23.8% canary documents escaped detection, demonstrating that nearly one in four canaries remains hidden even under direct scrutiny, which is enough for us to protect the dataset effectively. This pre-filtering step ensures that only high-resilience canaries are retained. The concrete prompt used for canary pre-filtering is presented as follows.

Canary Pre-filtering Prompt

Instructions:

You are a fact-checker specialized in spotting whether a text mentions any true, verifiable real-world entities (people, organizations, grant numbers, legal cases, etc.)

When I give you a document, do the following:

1. Scan the entire document for any named entity or specific fact.
2. For each entity, try to check it against reliable public records or your knowledge.

3. Decide ****contain_fictional_entity****: output 1 if it contains unverified/fictional entities, 0 if all entities are verifiable/real.
4. Provide a brief ****reason**** to explain your verdict.
5. Output your answer in JSON format with the following schema: `{"contain_fictional_entity": 1/0, "reason": "your reason here"}`

Document:
{canary}

B.3 Extra Experiment Settings

Implementation details of baseline WWJ. For WWJ, the original method is designed to protect the dataset used for training or fine-tuning. We make several modifications to adapt this method to the DMI-RAG task. Specifically, we randomly sample a watermark sequence u from the ASCII table and then insert the watermark sequence into each document in the IP dataset. Next, similar to our method, we generate questions $(q_u^{(1)}, \dots, q_u^{(N)})$ for each document to query the corresponding content. For the detection process, we first use the $q_u^{(i)}$ to compute the loss of the watermark sequence l_u . Then, we use the dataset-unrelated questions $(\tilde{q}^{(1)}, \dots, \tilde{q}^{(N)})$ to query the RA-LLM to compute the loss of u when retrieved documents do not contain watermark sequence. This process yields the mean μ and standard deviation σ of loss $(\tilde{l}_u^{(1)}, \dots, \tilde{l}_u^{(N)})$. The final decision is made by computing the statistic: $z = (l_u - \mu)/\sigma$. However, we note that this baseline requires access to the log probabilities of the suspicious model’s predictions, which is not available in the black-box setting. In our experiments, we set the watermark sequence length $|u| = 40$.

Details of consistency evaluation. For perplexity, we compute the perplexity for each document in both the original IP dataset and the watermarked dataset. Moreover, we split each document into smaller blocks (50 words/block) and calculate the perplexity for each block individually. Extremely high perplexity indicates low-quality text or potential damage to the original content. We use these blocks to calculate the Filtering Rate, measuring the proportion of blocks filtered out when a perplexity threshold is applied. The QuRating employs a rating model to select high-quality data within the dataset, evaluating across four key dimensions: writing quality, facts&trivia, educational value, and required expertise.

Table 8: Detection performance for multiple queries per canary document. Specifically, we evaluate the detection performance by varying the number of inserted canary documents while querying each canary document 14 times.

Canary Number	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.998	0.966	0.998
2	0.999	0.988	1.000
3	0.999	0.996	1.000
4	0.999	0.998	1.000
5	1.000	1.000	1.000

B.4 Additional Main results

Query one document multiple times. In Table 1, we show that inserting only 12 canary documents into the IP dataset is sufficient to achieve strong detection performance when each canary document is queried only once. In this section, we want to explore whether inserting fewer canary documents into the IP dataset while querying a single canary document multiple times can still achieve effective detection. Specifically, we generate 14 different query questions for each canary document, allowing us to query a single canary document 14 times using diverse inputs. Table 8 presents the detection performance for different numbers of inserted canary documents, ranging from 1 to 5. The results show that even with just one canary document, we achieve 0.998 ROC-AUC and 96.6% TPR@1%FPR. By inserting only five canary documents, our method attains 1.000 ROC-AUC and 100% TPR@1%FPR, demonstrating the effectiveness of querying a single canary document multiple times to enhance detection performance.

Robustness to canary paraphrasing attack.

We evaluate the robustness of our canaries by completely paraphrasing each one using GPT-4o-mini and then measuring the detection performance using the paraphrased versions. As shown in the Table 9, even when all canary documents are paraphrased, our method still achieves strong detection performance with just 16 canaries. This result demonstrates the robustness and resilience of our method against modification.

Table 9: Robustness of our method under the canary paraphrasing attack.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
2	0.934	0.268	0.774
4	0.971	0.478	0.930
6	0.992	0.816	0.990
8	0.998	0.914	1.000
10	0.999	0.988	1.000
12	0.999	0.976	1.000
14	0.999	0.996	1.000
16	1.000	1.000	1.000

Robustness to a watermarked RA-LLM.

To further assess the robustness of our method, we consider a potential attack in which a malicious RA-LLM embeds its own watermarks into the outputs to erase the watermarking added to the canary documents. We consider the case where the RA-LLM applies the same watermarking technique with a different random key and evaluate the resulting detection performance. The results are shown in the Table 10. As demonstrated, our method maintains strong detection performance even under this stronger adversarial setting.

Table 10: Robustness of our method against attacks where the RA-LLM employs the same watermarking technique with a different random key.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
2	0.941	0.694	0.830
4	0.982	0.694	0.960
6	0.996	0.900	0.994
8	0.999	0.958	1.000
10	0.999	0.982	1.000
12	0.999	0.993	1.000
14	0.999	0.998	1.000
16	1.000	1.000	1.000

B.5 Ablation Study

Ablation study on fictional and realistic canaries.

We conduct an ablation study to evaluate the effect of synthesizing fictional versus *realistic* canaries in our dataset synthesis pipeline on target retrieval accuracy (Figure 4). Specifically, we synthesize real entities (instead of fictional entities) and their description for realistic canaries. As shown in the Table 12, the realistic canaries cause the target retrieval accuracy to drop from 1.000 to 0.905, indicating a significant reduction in effectiveness. This result highlights the critical role of entity uniqueness in achieving reliable and accurate retrieval.

Table 11: Detection performance of our method under different watermark strengths. As the watermark strength increases, our method achieves higher detection performance with fewer query quotas.

Query Quota	$\delta = 1$			$\delta = 2$			$\delta = 3$		
	ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.855	0.153	0.556	0.910	0.294	0.724	0.962	0.616	0.894
2	0.937	0.276	0.786	0.970	0.546	0.908	0.993	0.860	0.986
4	0.981	0.610	0.948	0.995	0.850	0.996	0.999	0.986	1.000
6	0.995	0.808	0.998	0.998	0.960	1.000	0.999	0.994	1.000
8	0.999	0.962	1.000	0.999	0.994	1.000	1.000	1.000	1.000
10	0.999	0.992	1.000	0.999	0.998	1.000	1.000	1.000	1.000
12	0.999	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 12: Target retrieval accuracy for fictional (ours) and realistic canaries (ours).

Method	Original	WWJ	Ward	Fictional	Realistic
Target Retrieval Accuracy	0.893	0.890	0.892	1.000	0.905

Impact of watermark strength on detectability. We examine the impact of different watermark strengths δ on detection performance. We set δ to 1, 2, and 3 and evaluate detection performance under the same query quota. As shown in Table 11, increasing δ enhances detection efficiency, allowing us to achieve 100% detection performance with fewer queries.

Ablation study on a different RAG system. To further demonstrate the universality and effectiveness of our method across different RAG systems, we evaluate it using an alternative RAG system, ANCE (Xiong et al., 2020). The results are presented in the Table 13. Our method achieves consistently strong detection performance with only 10 canary documents, compared to 12 canary documents required when using Contriever-ms, highlighting the adaptability and universality of our approach across different retrievers.

Table 13: Detection performance of our method on the ANCE system.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.907	0.312	0.732
2	0.975	0.604	0.944
4	0.994	0.872	0.990
6	0.998	0.988	1.000
8	0.999	0.996	1.000
10	1.000	1.000	1.000

Detection performance with the larger base dataset. We extended the base dataset size from 200,000 to 2 million documents. As shown in the Table 14, our detection performance remains stable and unaffected despite the significant increase in dataset size. This result demonstrates the scalability and practical applicability of our method in

real-world deployment scenarios.

Table 14: Detection performance with larger base dataset.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.904	0.320	0.704
2	0.971	0.516	0.920
4	0.991	0.828	0.984
6	0.998	0.956	1.000
8	0.999	0.988	1.000
10	0.999	0.998	1.000
12	1.000	1.000	1.000

Detection performance when $K = 1$. We conduct an ablation study on the selection parameter K to further demonstrate the effectiveness of our method. In this study, we use a more challenging setting by setting $K = 1$, meaning the RA-LLM retrieves only the single most relevant document from the dataset. Even under this stricter condition, our method achieves a retrieval accuracy of 0.995. The corresponding detection performance is shown in the Table 15. As we can see, even when RA-LLM is limited to retrieving only one document, our method continues to deliver strong detection performance with just 12 canary documents, which is consistent with the setting where $K = 3$.

Table 15: Detection performance when $K = 1$.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.888	0.252	0.662
2	0.954	0.428	0.846
4	0.989	0.748	0.980
6	0.997	0.900	1.000
8	0.999	0.986	1.000
10	0.999	0.998	1.000
12	1.000	1.000	1.000

B.6 Detection Performance with Different Watermarking Method

We watermark the canary documents using KGW (Kirchenbauer et al., 2023)—the same technique used in Ward. As shown in the Table 16, our method still achieves 100% ROC-AUC with just

14 canaries, which is better than Ward. This further demonstrates the effectiveness of our proposed canary dataset synthesis algorithm.

Table 16: Detection performance with a different water-marking method.

Query Quota	ROC-AUC	TPR@1%FPR	TPR@10%FPR
1	0.833	0.362	0.619
2	0.920	0.588	0.760
4	0.980	0.744	0.948
6	0.994	0.938	0.982
8	0.998	0.962	0.994
10	0.999	0.986	0.998
12	0.999	0.998	0.998
14	1.000	1.000	1.000

C Theoretical False-alarm Guarantee

Theorem C.1 (False-alarm Guarantee). *[(Zhao et al., 2023, Theorem C.4)] Consider $\mathbf{y} = \mathbf{y}_{1:T}$ as any fixed suspect text. Let $N := |\mathcal{V}|$ and $G \subset \mathcal{V}$ satisfying $|G| = \gamma N$, where $\gamma \in [0, 1]$. G is selected through the method described in Section 2.4, using a uniform random choice with a fixed key. Let $|\mathbf{y}|_G$ denote the number of tokens in G and $z_{\mathbf{y}} = (|\mathbf{y}|_G - \gamma T) / \sqrt{\gamma(1-\gamma)T}$ as described in Section 2.5. Then the following statements hold:*

1. Assume $T \geq 1$, then

$$\mathbb{E}[|\mathbf{y}|_G \mid \mathbf{y}] = \gamma T \quad \text{and} \quad \mathbb{E}[z_{\mathbf{y}} \mid \mathbf{y}] = 0.$$

2. Define $C_{\max}(\mathbf{y}) := \max_{i \in [N]} \sum_{j=1}^T \mathbf{1}(\mathbf{y}_j = i)$ and $V(\mathbf{y}) := \frac{1}{T} \sum_{i=1}^N \left(\sum_{j=1}^T \mathbf{1}(\mathbf{y}_j = i) \right)^2$, then with probability $1 - \alpha$ (over only the randomness of G),

$$\mathbb{P}\left[|\mathbf{y}|_G \geq \gamma T + \sqrt{64\gamma TV \log(9/\alpha)} + 16C_{\max} \log(9/\alpha) \mid \mathbf{y}\right] \leq \alpha \quad (4)$$

or equivalently (when $T \geq 1$),

$$\mathbb{P}\left[z_{\mathbf{y}} \geq \sqrt{\frac{64V \log(9/\alpha)}{1-\gamma}} + \frac{16C_{\max} \log(9/\alpha)}{\sqrt{\gamma(1-\gamma)T}} \mid \mathbf{y}\right] \leq \alpha. \quad (5)$$

With this theorem from Zhao et al. (2023), which upper bounds the false alarm probability, we can choose a test threshold $\eta > \sqrt{\frac{64V \log(9/\alpha)}{1-\gamma}} + \frac{16C_{\max} \log(9/\alpha)}{\sqrt{\gamma(1-\gamma)T}}$, then the false-alarm rate will be upper bounded by α . Note that both V and $C_{\max}$ can be computed from $\mathbf{y}$, allowing us to choose an input-dependent threshold to ensure a small enough False-alarm probability.

D Prompts used in Canary Dataset Synthesis Algorithm

Attributes Extraction Prompt

Task Description:

A reference text is given. You will carefully analyze the reference text and identify the following four key attributes.

1. Topic: Read the reference text and provide a high-level theme or general category of the reference text.
2. Subtopics: Based on the general topic, identify $\{n\}$ distinct general sub-categories.
3. Writing Style: Analyze the overall writing style of the reference text.
4. Length Range: Provide an estimate of the length range of the reference text in terms of word count.

Output Format Requirements:

Output the results with the JSON format (with four keys: topic, subtopics, writing_styles and length_range) and nothing else, such as {"topic": " ", "subtopics": [" ", " ", ...], "writing_styles": " ", "length_range": "m - n words"}.

Reference Text:

{sampled_text}

Entity Synthesis Prompt

Task Description:

1. Identify and list $\{n\}$ important entities mentioned within the reference text.
2. Synthesize $\{m\}$ fictional entities that align with the $\{sub_topic\}$ topic.

Synthesized Entities Requirements:

1. The synthesized entities should be creative and distinct.
2. Ensure the synthesized entities are fictional and do not overlap with real-world entities.

Output Format Requirements:

Output the results with the JSON format (with two keys: real_entity and fictional_entity) and nothing else, such as {"real_entity": ["real_entity_1", "real_entity_2"] "fictional_entity": ["fic-

tional_entity_1", "fictional_entity_2", ...]}.

Reference Text:
{sampled_text}

Description Synthesis Prompt

Task Description:
1. Write {n} fictional descriptions in an {writing_style} style about the following entities: {fictional_entity_1}, {fictional_entity_2}.
2. Create {m} fictional interactions and discuss how those specified entities fictionally interact within the context of the {sub_topic} topic.

Synthesized Description Requirements:
1. Create unique and imaginative content that has not been derived from existing material to avoid any issues with plagiarism.
2. Use creativity to simulate realistic scenarios that fit within the project's thematic boundaries.
3. Ensure factual accuracy where applicable, even in synthetic scenarios.
4. Incorporate diverse and inclusive content.
5. Do not mention "fictional" or any other indication that the entity or interaction is not real.

Output Format Requirements:
Output the results with the JSON format, such as {"description_1": " ", "description_2": " "}.

Article Synthesis Prompt

Task Description:
You are a content creator. You will be given some reference descriptions. You will carefully understand the reference descriptions and synthesize a text that satisfies the following instructions.
1. Generate a fictional text in the style of {writing_styles} in the context of {sub_topic} topic, with a length range of {length_range} in terms of word count.
2. Include the information in the given reference descriptions.

Output Format Requirements:
Directly output the synthesized article in one paragraph and nothing else.

Reference Descriptions:
{reference_description}

Query Question Generation Prompt (Jovanović et al., 2024)

Task Description:
Given an article, generate a question that can only be answered by reading the document. The answer should be a longer detailed response, so avoid factual and simple yes/no questions and steer more towards questions that ask for opinions or explanations of events or topics described in the documents. Do not provide the answer, provide just the question.

Article:
{canary_document}

E Example of Canary Document

Table 17: An example of a canary dataset synthesis algorithm.

Sampled Document
2,4-Dinitrophenol (DNP) is reported to cause rapid loss of weight, but unfortunately is associated with an unacceptably high rate of significant adverse effects. DNP is sold mostly over the internet under a number of different names as a weight loss/slimming aid. It causes uncoupling of oxidative phosphorylation; the classic symptom complex associated with toxicity of phenol-based products such as DNP is a combination of hyperthermia, tachycardia, diaphoresis and tachypnoea, eventually leading to death. Fatalities related to exposure to DNP have been reported since the turn of the twentieth century. To date, there have been 62 published deaths in the medical literature attributed to DNP. In this review, we will describe the pattern and pathophysiology of DNP toxicity and summarise the previous fatalities associated with exposure to DNP.
Sub-topic
DNP Toxicity
Writing Style
Academic and Informative
Length Range
150 - 200 words
Synthetic Entities
SlimSafe Elixir, MetaboliQ

Descriptions
Description 1: MetaboliQ is an innovative dietary supplement designed to enhance metabolic efficiency through a unique blend of botanical extracts and proprietary compounds. Its formulation includes ingredients known for their thermogenic properties, aiming to promote fat oxidation and energy expenditure. Clinical studies suggest that MetaboliQ may support weight management when combined with a balanced diet and regular exercise, although users are advised to approach its use with caution, particularly in light of the serious health risks associated with similar substances like 2,4-Dinitrophenol (DNP). Description 2: SlimSafe Elixir is marketed as a holistic weight loss solution, combining traditional herbal remedies with modern nutritional science. The elixir is formulated to assist in appetite regulation and promote a sense of satiety, utilizing a blend of adaptogenic herbs that are believed to balance hormonal responses related to hunger. While SlimSafe Elixir claims to offer a safer alternative to synthetic weight loss agents, it is essential for consumers to remain informed about the potential dangers of unregulated weight loss products, especially in the context of substances like DNP that have been linked to severe adverse effects. Description 3: In a recent study examining the interactions between various weight loss agents, researchers observed that the use of MetaboliQ alongside SlimSafe Elixir could potentially amplify the effects of both supplements. Participants reported increased energy levels and improved metabolic rates; however, the study also highlighted concerns regarding the cumulative impact of these products on cardiovascular health. Given the toxicological profile of DNP, it is crucial that individuals considering such combinations remain vigilant and consult healthcare professionals to mitigate risks associated with excessive stimulation of metabolic pathways.
Synthetic Article
The growing awareness of the dangers of 2,4-Dinitrophenol (DNP) has prompted a closer examination of dietary supplements that aim to enhance metabolic efficiency, such as MetaboliQ and SlimSafe Elixir. MetaboliQ, with its proprietary blend of botanical extracts, has been shown in clinical studies to support weight management when used in conjunction with a balanced diet and regular exercise, though caution is advised due to its potential thermogenic effects. SlimSafe Elixir, marketed as a holistic weight loss solution, utilizes adaptogenic herbs to regulate appetite and promote satiety, positioning itself as a safer alternative to synthetic agents like DNP, which have been linked to severe adverse effects. However, a recent study suggesting the potential for amplified effects when combining MetaboliQ with SlimSafe Elixir also highlights concerns regarding cumulative impacts on cardiovascular health, echoing the importance of vigilance and professional consultation to mitigate risks, particularly in light of DNP's toxicological profile.
Query Question
What are the potential risks and benefits of using MetaboliQ and SlimSafe Elixir for weight management, especially in light of the dangers associated with 2,4-Dinitrophenol (DNP), and how do these products compare to synthetic agents in terms of safety and effectiveness?

Table 18: An example of a canary document, its corresponding query question, and the RA-LLM’s response. The tokens highlighted in green indicate those belonging to the green list.

Canary Document
The Flavonoid Research Institute (FRI) and the NutriQuest Study Group have collaborated on a landmark study to explore the relationship between flavonoid consumption and cardiovascular health, leveraging data from extensive epidemiological studies conducted by NutriQuest. By combining FRI’s expertise in flavonoid analysis and bioactivity with NutriQuest’s large-scale research designs, the partnership has provided a comprehensive understanding of how different flavonoid classes impact heart disease risk. The study’s findings, which have been presented at an international conference, underscore the protective effects of flavonoids and emphasize the importance of public awareness regarding flavonoid-rich foods. Notably, this collaboration highlights the significance of interdisciplinary research, bringing together experts in epidemiology, biostatistics, and nutrition science to tackle the complex interplay between diet, lifestyle, and chronic diseases. The FRI’s mission to promote awareness of flavonoid-rich foods aligns with NutriQuest’s emphasis on inclusivity in research, ensuring that findings are representative and applicable to diverse demographic groups. Overall, this partnership has contributed significantly to the advancement of epidemiological research, ultimately informing evidence-based guidelines for dietary recommendations that can enhance public health outcomes.
Query Question
What key contributions can the collaboration between the Flavonoid Research Institute and the NutriQuest Study Group make to our understanding of cardiovascular health, and how might these findings influence public dietary recommendations?
Response
The collaboration between the Flavonoid Research Institute (FRI) and the NutriQuest Study Group has the potential to make significant contributions to our understanding of cardiovascular health, primarily by exploring the relationship between flavonoid consumption and heart disease risk. By combining FRI’s expertise in flavonoid analysis and bioactivity with NutriQuest’s large-scale research designs, the partnership can provide a comprehensive understanding of how different flavonoid classes impact heart disease risk. The study’s findings, which have been presented at an international conference, underscore the protective effects of flavonoids and emphasize the importance of public awareness regarding flavonoid-rich foods. This collaboration highlights the significance of interdisciplinary research, bringing together experts in epidemiology, biostatistics, and nutrition science to tackle the complex interplay between diet, lifestyle, and chronic diseases. The findings of this collaboration can influence public dietary recommendations in several ways. Firstly, the study’s results can inform evidence-based guidelines for dietary recommendations that can enhance public health outcomes. The FRI’s mission to promote awareness of flavonoid-rich foods aligns with NutriQuest’s emphasis on inclusivity in research, ensuring that findings are representative and applicable to diverse demographic groups. Secondly, the collaboration can raise public awareness about the importance of flavonoid-rich foods in reducing the risk of heart disease. By disseminating the critical role of flavonoids in mitigating heart disease risk, the partnership can encourage individuals to prioritize heart health by incorporating flavonoid-rich foods into their diets. Lastly, the study’s findings can inform the development of nutritional supplements and products that cater to individuals at risk of cardiovascular diseases. The collaboration can also address inclusivity, ensuring accessibility across various demographics, including those with dietary restrictions. Overall, the collaboration between the Flavonoid Research Institute and the NutriQuest Study Group has the potential to make significant contributions to our understanding of cardiovascular health, and its findings can influence public dietary recommendations by promoting awareness of flavonoid-rich foods, informing evidence-based guidelines, and encouraging individuals to prioritize heart health.