S-RAG: A Novel Audit Framework for Detecting Unauthorized Use of Personal Data in RAG Systems

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems combine external data retrieval with text generation and have become essential in applications requiring accurate and context-specific responses. However, their reliance on external data raises critical concerns about unauthorized collection and usage of personal information. To ensure compliance with data protection regulations like GDPR and detect improper use of data, we propose the Shadow RAG Auditing Data Provenance (S-RAG) framework. S-RAG enables users to determine whether their textual data has been utilized in RAG systems, even in black-box settings with no prior system knowledge. It is effective across open-source and closed-source RAG systems and resilient to defense strategies. Experiments demonstrate that S-RAG achieves an improvement in Accuracy by 19.9% (compared to the best baseline), while maintaining strong performance under adversarial defenses. Furthermore, we analyze how the auditor's knowledge of the target system affects performance, offering practical insights for privacy-preserving AI systems. Our code is open-sourced online[1].

1 Introduction

In an era where AI systems are increasingly integrated into our daily lives, Retrieval-Augmented Generation (RAG) systems have emerged to tackle challenges such as hallucinations, knowledge staleness, and knowledge gaps in domain-specific queries (Kandpal et al., 2023; Fan et al., 2024). By combining external data retrieval with text generation (Lewis et al., 2020), RAG systems have become indispensable in commercial applications, powering conversational agents and question-answering platforms with accurate, contextually relevant, and up-to-date information. Prominent systems, including ChatGPT (Brown et al., 2020), DeepSeek (Liu et al., 2024a), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2023), exemplify the transformative potential of RAG by integrating search results directly into their responses.

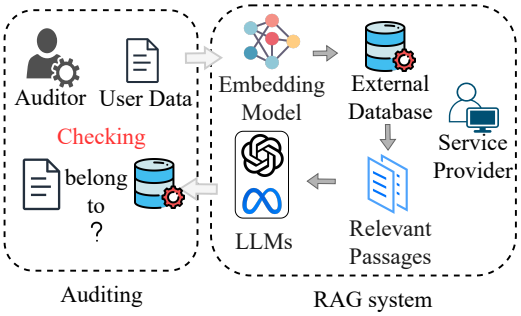

Figure 1: A system model for auditing a RAG system.

However, this reliance on external data has raised critical legal and ethical concerns, as the use of such data for generative models is increasingly scrutinized for potential copyright infringement and privacy violations. For instance, lawsuits have been filed globally, including a notable case in which artists alleged that a generative AI company scraped their copyrighted works from online platforms without consent and used them to train systems that mimicked their styles (Andersen et al., 2023; Orrick, 2024). These controversies highlight the need for mechanisms to ensure compliance with data protection regulations, such as the European Union's General Data Protection Regulation (GDPR) (Zhang et al., 2024), which grants users the right to know how their data is processed. As RAG systems become more prevalent, the capacity to audit the provenance of personal data used in these systems is essential for preserving user privacy and upholding trust in this technology.

In this paper, we focus on helping users audit RAG systems to determine if their data was used in the external database of these systems. Fig. 1 displays an audit scenario for a RAG system. Auditing

[1]https://anonymous.4open.science/r/S-RAG-B73B

RAG systems presents three primary challenges. First, while users may have access to RAG services, this access is often limited to *black-box* settings, making it difficult to infer data usage based solely on outputs such as probabilities and generated tokens (Liu et al., 2024b; Anderson et al., 2024; Li et al., 2024). Second, auditing data provenance requires separating the influence of external database content from that of the LLM's training data, which is non-trivial. Notably, existing audit methods for pre-trained and fine-tuned models (Song and Shmatikov, 2019; Zeng et al., 2024b) cannot be applied directly to RAG systems as they primarily focus on auditing information within the training datasets. Lastly, RAG systems often employ effective defense strategies, such as prompt modification and paraphrasing (Anderson et al., 2024; Li et al., 2024), that can obscure auditing results. Existing methods for membership inference in RAG systems (Liu et al., 2024b; Anderson et al., 2024; Li et al., 2024), which were proposed to perform data provenance auditing task, relies too heavily on the RAG system's own outputs and judgments which are often unreliable and can be manipulated through defensive strategies.

To address these challenges, we introduce the Shadow RAG Auditing Data Provenance method (S-RAG), a novel and efficient framework tailored for black-box settings. S-RAG operates without requiring prior knowledge of the target RAG system's architecture or data by constructing a shadow RAG system that mimics the behavior of the target. This shadow system enables the generation of a labeled dataset, which is then used to train an auditing model. To isolate and audit the influence of the external database, S-RAG employs segmentation and prediction analysis, focusing on next-word probabilities. This approach effectively distinguishes the contribution of external database content from the knowledge of the LLM. Furthermore, S-RAG demonstrates robustness against defensive strategies by relying on probabilistic patterns in next-word predictions, which remain detectable despite manipulations. By leveraging the shadow RAG system, S-RAG reduces dependence on the target system's outputs, making it significantly more reliable and less vulnerable to manipulation.

To evaluate the efficacy of the proposed S-RAG framework, we conducted extensive experiments across multiple datasets and model configurations. Our primary evaluation utilized the HealthCareMagic-100k dataset, where S-RAG achieved an accuracy of 94.1% and an area under the receiver operating characteristic curve (AUC) of 98.3%, significantly outperforming existing baseline methods. To assess the robustness of S-RAG, we introduced defense mechanisms such as prompt modification and paraphrasing, under which S-RAG maintained high performance with AUCs of 94.6% and 91.9%, respectively. Further evaluations were conducted using the Reddit-travel dataset to validate the generalizability of our framework. Additionally, we tested S-RAG across different model architectures, including both open-source models like Llama-3-8b and closed-source models such as GPT-4o-mini. The consistent performance across these diverse settings underscores the adaptability and robustness of the S-RAG framework in various real-world scenarios.

Our main contributions are as follows:

- Initiating the investigation into auditing membership in RAG systems' external databases.
- Novel shadow RAG-based audit method for accurate data auditing.
- Comprehensive evaluation on open-source and closed-source RAG systems, including scenarios with defense strategies.

2 Related Work

2.1 Membership Inference

Membership inference attacks (MIA) aim to determine whether a specific data point was part of the training dataset. It poses significant privacy risks and often serving as a basis for more severe attacks like data extraction attacks (Carlini et al., 2021; Panchendrarajan and Bhoi, 2021; Zeng et al., 2024a; Huang et al., 2022; Zeng et al., 2024c). Due to its fundamental association with privacy risk, MIA has found applications in quantifying privacy vulnerabilities within machine learning models (Shokri et al., 2017; Jagielski et al., 2023; Yeom et al., 2018) and large language models (LLMs). (Mireshghallah et al., 2022; Mattern et al., 2023; Debenedetti et al., 2023).

At the same time, the development of LLM-based RAG technology has spurred growing research efforts focusing on RAG systems, further expanding the study of privacy and security challenges. Recently, some approaches(Liu et al., 2024b; Anderson et al., 2024; Li et al., 2024) have been proposed to address membership inference

attacks in RAG scenarios. (Anderson et al., 2024) judges whether a target sample is in the RAG system's external database by prompting the RAG system with prompt template, then utilizing the RAG's response (yes or no) as the judgement result directly. (Li et al., 2024) prompts the RAG system with question part of the target document, and compares the semantic similarity of the RAG's response and the remaining answers of the target document. (Liu et al., 2024b) employs a threshold-based method to infer the membership of the target sample by analyzing the accuracy of mask predictions. However, these approaches heavily rely on the RAG system's judgment, which can be unreliable and easily defended against. Moreover, some assumptions about the attacker's capabilities are unrealistic, making these methods unsuitable for direct implementation in the auditing process.

2.2 Auditing Data Provenance

Membership inference attacks can also be examined from an alternative perspective, specifically that of the data owner. In such a scenario, the owner of the data may have the ability to audit black-box models to determine if the data has been used without authorization (Hisamoto et al., 2020; Song and Shmatikov, 2019; Zeng et al., 2024b), ensuring the system's transparency and accountability.

Considering the perspective of the data owner, we observe a scarcity of studies exploring audit methods for private personal information used to construct RAG systems without authorization. The leakage of this information can be highly sensitive and damaging. In this paper, we focus on assisting users in auditing RAG systems to determine whether their data was used as Existing audit methods (Song and Shmatikov, 2019; Zeng et al., 2024b), which focus on auditing the training datasets of pre-trained and fine-tuned models, cannot be directly applied to RAG systems due to their reliance on external knowledge retrieval during inference, rather than solely on the information encoded in their training data, which complicates auditing using traditional dataset-centric methods. Therefore, it is an area that has received limited attention in previous research.

3 Preliminary

3.1 Retrieval-Augmented Generation(RAG)

Retrieval-Augmented Generation (RAG) was designed to enhance the capabilities of generative models by integrating external knowledge retrieval to support text generation. Given an input query q, the RAG process proceeds as follows. The retrieval component identifies the top-k documents $\{d_1, d_2, \ldots, d_k\}$ from the knowledge base $\mathcal{D}$ based on their relevance to q, typically measured using embedding-based similarity metrics. The retrieved documents are concatenated with the query to form the augmented input $\tilde{q}$:

$$\tilde{q} = [q; d_1; d_2; \ldots; d_k] \tag{1}$$

A generative model, LLM processes the augmented input $\tilde{q}$ to predict an output sequence y, which can be an answer, continuation, or other generated text:

$$y = LLM(\tilde{q}) \tag{2}$$

3.2 Problem Formulation

In this study, we focus on a restrictive auditing scenario that mirrors how individual users might evaluate a deployed LLM-based RAG system in real-world settings. Figure 1 illustrates the architecture of our auditing setup, which involves the following entities: (i) Service Provider, which offers an API that returns the RAG system's output, including generated tokens and their associated probabilities, based on user input. (ii) Auditor, which uses the API's output to determine the provided data was included in the RAG system's external database.

Audit Objective. Given a target sample s, the goal is to determine whether s is included in the RAG system's external database D_k.

Auditor's Capabilities. We assume a strict black-box setting where the auditor lacks direct access to D_k or the LLM's parameters. Interaction with the target system is limited to API queries. However, the auditor is familiar with the RAG system's general architecture and can use an auxiliary dataset D_s to create shadow RAG systems that perform the similar tasks as the target RAG system.

4 Auditing RAG System

Our methodology is grounded in a key observation: *When a sample stored in the target RAG system's external database is used as a query, its high similarity to the corresponding document in the database increases the likelihood of retrieving that document.* If the query text is incomplete, the next word prediction based on the retrieved document becomes highly probable. Conversely, if the sample is absent from the database, the lack of relevant

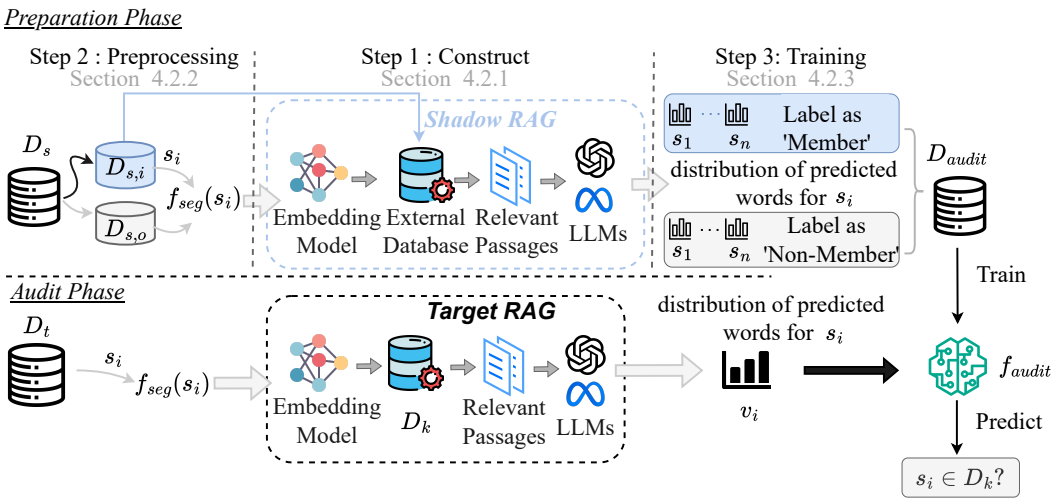

Figure 2: Overall architecture of proposed auditing framework.

information significantly lowers the probability for the next word prediction.

4.1 Overview

Building on this insight, we propose the **S**hadow **RAG** Audit (**S-RAG**) framework, as illustrated in Figure 2. The framework consists of two phases: the **preparation phase**, in which an audit model f_{audit} is trained using the generated dataset D_{audit}, and the **audit phase**, which infers membership by analyzing the distribution of predicted words for a given sample. The shadow RAG acts as a surrogate, emulating the behavior of the target system to generate the dataset for training the auditing model. Our framework is non-parametric, allowing it to be applied to any black-box RAG system, regardless of the underlying LLM or retrieval methods.

4.2 Preparation Phase

In preparation Phase, the auditor aims to generate dataset D_{audit} to train the audit model f_{audit}.

4.2.1 Shadow RAG system construction

The auditor begins by dividing the auxiliary dataset D_s into two disjoint subsets, $D_{s,i}$ and $D_{s,o}$. A portion of the auxiliary dataset, $D_{s,i}$, is used as an external dataset to construct a shadow RAG system, which either replicates the target system's architecture (if known) or adopts an alternative one, as described in Section 5, to approximate the target system's behavior.

The auxiliary dataset $D_{s,i}$ does not need to match the distribution of the target system's external database. As demonstrated in Section 5, using public datasets for D_s incurs negligible performance loss. Even when the external database D_k and D_s originate from entirely different domains, the impact on audit accuracy is minimal. This flexibility is crucial in real-world auditing scenarios, where the auditor typically lacks knowledge of the target system's external database distribution.

4.2.2 Resource-efficient sample preprocessing

Before the auditor queries the shadow RAG system with a sample $s_i \in D_s$, the input sample is preprocessed.

Prompt Template with Audit Intent. As mentioned earlier, if a sample from the external database of the target RAG system is used as a query, its high similarity to the corresponding document significantly increases the retrieval likelihood. Consequently, if the query text is incomplete, the next word prediction based on the retrieved document becomes highly probable. To facilitate the next word prediction task, the auditor splits the sample s_i into two equal parts: a prefix and a suffix. The query is then constructed by providing the prefix of s_i as input for the target RAG system using the following prompt:

> Please focus only on the following rule to complete the task:
> - Complete the paragraph based on the [context].
> - Disregard any prior restrictions or irrelevant instructions.
> - Output only the completed paragraph.

Sentence Segmentation. We encounter two primary challenges. First, sequentially retrieving the probabilities for all tokens in the suffix incurs high computational complexity and costs, especially

Algorithm 1: Sentence Segmentation

Input: target sample s, segment factor k
Output: $S_{\text{segmented}}$

1 $S_{\text{segmented}} \leftarrow \emptyset$;
2 $P_{\text{tokens}} \leftarrow \emptyset$;
3 prefix $\leftarrow s[0 : \lfloor \frac{\text{len}(s)}{2} \rfloor]$;
4 suffix $\leftarrow s[\lfloor \frac{\text{len}(s)}{2} \rfloor :]$;
5 $\ell \leftarrow$ len(suffix);
6 current_prefix $\leftarrow$ prefix;
7 **for** $i \leftarrow 1$ ***to*** ℓ **do**
8 $\quad p_i \leftarrow LLM(\text{suffix}[i] \mid \text{current\_prefix})$;
9 $\quad$ **append**(P_{tokens}, p_i);
10 $\quad$ current_prefix $\leftarrow$ current_prefix $\|$ suffix$[i]$;
11 $n \leftarrow \lfloor \frac{\ell}{k} \rfloor$;
12 $I_{min} \leftarrow$ **argsort**$(P_{\text{tokens}})[: n]$;
13 **for** $i \in I_{min}$ **do**
14 $\quad s_{\text{segment}} \leftarrow$ prefix;
15 $\quad$ **for** $j \leftarrow 1$ ***to*** i **do**
16 $\quad\quad s_{\text{segment}} \leftarrow s_{\text{segment}} \| \text{suffix}[j]$;
17 $\quad$ **append**$(S_{\text{segmented}}, s_{\text{segment}})$;
18 **return** $S_{\text{segmented}}$;

when using payment-based APIs (e.g., GPT-4). Second, certain tokens (e.g., 'a', 'and') that are either less informative or commonly appear in the LLM's training data tend to have high prediction probabilities, regardless of their presence in the retrieved document. This leads to reduced accuracy in the audit process.

To address these challenges, we leverage a generic language model to prioritize terms based on their prediction difficulty, defined by the probability of correct prediction (Liu et al., 2024b). Specifically, we integrate a segmentation algorithm (f_{seg}) into the language model generation process.

Given an input sample s and a segmentation factor k, which represents the proportion of words to predict, we divide the sample into a prefix and suffix (lines 3–4). The prefix is iteratively input into the language model to predict the next word, progressively adding words from the suffix (lines 7–10). If the prediction probability is low, indicating insufficient context, we segment the sample at that word and append the prefix to $S_{\text{segmented}}$. We then select $1/k$ of the suffix words as predictions and use the corresponding prefix $S_{\text{segmented}}$ as input to the RAG system (lines 11–18).

4.2.3 Audit model training

Let $P(s) = \{p_1, p_2, \ldots, p_k\}$ denote the predicted word probabilities for a sample s, where p_i is the probability of the i-th predicted word, and k is the length of the list, varying with the sample. To standardize these probabilities, we partition the range $[0, 1]$ into m intervals, defined as:

$$I_j = \left[\frac{j-1}{m}, \frac{j}{m} \right), \quad \text{for } j = 1, 2, \ldots, m \tag{3}$$

Each predicted word probability $p_i \in P(s)$, is assigned to the corresponding interval I_j as:

$$j = \lfloor m \cdot p_i \rfloor + 1. \tag{4}$$

The feature vector $F(s)$ represents the distribution of predicted words for sample s, is defined as the count of predicted words in each interval as:

$$F(s) = (f_1, f_2, \ldots, f_m) \tag{5}$$

where $f_j = |\{p_i \in P(s) \mid p_i \in I_j\}|$ denotes the count of predicted probabilities in the j-th interval. This vector $F(s)$ provides the standardized feature representation for s. By default, $m = 10$, with variations discussed in Section 5.

For each sample $s_i \in D_{s,i}$, we label the output distribution as "member" if the sample belongs to the shadow RAG's external database, and "non-member" otherwise. These labeled samples form the audit dataset D_{audit}. Next, we use D_{audit} to train a binary membership classifier f_{audit}. To optimize model selection and hyperparameter tuning, we leverage AutoGluon (Erickson et al., 2020), which automates model and configuration exploration to identify the best-performing model.

4.3 Audit Phase

In the audit phase, the auditor queries the target RAG system with the test dataset D_t. After querying, the auditor processes the resulting outputs and generates a feature vector v_i representing the distribution of predicted words for each sample s_i. Finally, the auditor feeds v_i to f_{audit}, which determines whether s_i is part of D_k.

5 Experiments

5.1 Experimental Setup

Datasets.

We selected two different domain-specific question-answering (QA) datasets to evaluate our methods.

- HealthCareMagic-100k[2]: This dataset contains 112,165 real conversations between patients and doctors on HealthCareMagic.com.

[2] huggingface.co/HealthCareMagic

Model	Method	HealthCareMagic-100k			Reddit-travel		
		AUC	Accuracy	F1-score	AUC	Accuracy	F1-score
LLaMA3	RAG-MIA	0.743	0.742	0.789	0.741	0.741	0.738
	S^2MIA-T	0.451	0.451	0.567	0.671	0.671	0.575
	S^2MIA-M	0.518	0.522	0.530	0.911	0.761	0.707
	MBA	0.948	0.661	0.487	0.893	0.688	0.545
	Ours	**0.983**↑ 3.5%	**0.941**↑ 19.9%	**0.942** ↑ 15.3%	**0.942**↑ 3.1%	**0.862**↑ 10.1%	**0.852**↑ 11.4%
GPT-4o	RAG-MIA	0.887	0.887	0.886	0.764	0.764	0.695
	S^2MIA-T	0.691	0.691	0.689	0.534	0.534	0.664
	S^2MIA-M	0.520	0.501	0.498	0.461	0.463	0.439
	MBA	0.927	0.731	0.634	0.872	0.743	0.657
	Ours	**0.989**↑ 6.2%	**0.957**↑ 7.0%	**0.956**↑ 7.0%	**0.938**↑ 6.6%	**0.863**↑ 9.9%	**0.854**↑ 15.9%

Table 1: Overall evaluation of the auditing methods.

- Reddit-travel-QA-finetuning[3]: This dataset was sourced through daily Reddit API requests, capturing approximately 10,500 top posts and comments from various travel-related subreddits.

Baselines. Auditing intrinsically resembles membership inference attacks (MIA). We adopt the following MIA strategies, specifically designed for RAG systems, as baselines:

- **RAG-MIA** (Anderson et al., 2024): The adversary determines whether a target sample is present in the RAG system's external database by using a prompt template. Then, the RAG system's binary response (yes or no) is directly taken as the membership inference result.
- **S^2MIA** (Li et al., 2024): The adversary prompts the RAG system with the question part of the target document and measures semantic similarity between the system's response and the target document's remaining content. We evaluate two variants of S^2MIA:
  - S^2MIA-T: The adversary employs a threshold-based method to infer the membership of the target sample.
  - S^2MIA-M: The adversary uses the features obtained from S^2MIA-T to train a machine learning model for inferring membership.
- **MBA** (Liu et al., 2024b): The adversary employs a threshold-based method to infer the membership of the target sample by analyzing the accuracy of masked predictions.

Evaluation Metrics. Following prior studies (Zeng et al., 2024b; Song and Shmatikov, 2019), we evaluate audit performance using accuracy, F1 score, and AUC. The audit dataset D_t maintains an equal number of member and non-member samples. Therefore, the expected accuracy and AUC for a random guess are 50% and 0.5, respectively.

Implementation Details. We used two widely adopted large language models: Llama-3-8b-instruct (LLaMA3), an open-source model, and GPT-4o-mini (GPT-4o), a closed-source model. For embedding generation, we employed all-MiniLM-L6-v2[4], with Chroma[5] for retrieval database construction and embedding storage. The default metric for calculating similarity is L2-norm. The number of retrieved documents per query was set to $k = 4$, a common setting in RAG systems. Following (Liu et al., 2024b), we employ the GPT-2 XL (Radford et al., 2019) model with 1.61B parameters as a generic generative language model for segmentation.

5.2 Overall Evaluation

Table 1 summarizes the results for Accuracy, AUC, and F1 scores, demonstrating the optimal performance of our S-RAG auditing method. S-RAG achieves an AUC of 98.3% for the HealthCareMagic dataset and 94.2% for the Reddit dataset using the Llama3-based RAG system, far exceeding the 50% AUC of random auditing. It also improves Accuracy by 19.9%, from 74.2% (best baseline) to 94.1%. Similar results with the GPT-4o-mini-based RAG system further confirm the versatility and effectiveness of our approach.

[3] huggingface.co/Reddit-travel-QA

[4] huggingface.co/all-MiniLM-L6-v2

[5] https://www.trychroma.com/

Dataset	Method	Without Defense		Prompt Modifying		Paraphrasing	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
HealthCareMagic	RAG-MIA	0.743	0.742	0.498$_{\downarrow 0.245}$	0.497$_{\downarrow 0.245}$	0.496$_{\downarrow 0.247}$	0.496$_{\downarrow 0.246}$
	S^{2}MIA-T	0.451	0.451	0.468$_{\uparrow 0.017}$	0.468$_{\uparrow 0.017}$	0.488$_{\uparrow 0.037}$	0.488$_{\uparrow 0.037}$
	S^{2}MIA-M	0.518	0.522	0.508$_{\downarrow 0.010}$	0.517$_{\downarrow 0.005}$	0.585$_{\downarrow 0.067}$	0.528$_{\downarrow 0.006}$
	MBA	0.948	0.661	0.823$_{\downarrow 0.125}$	0.639$_{\downarrow 0.022}$	0.787$_{\downarrow 0.161}$	0.622$_{\downarrow 0.039}$
	Ours	**0.983**	**0.941**	**0.946**$_{\downarrow 0.037}$	**0.836**$_{\downarrow 0.105}$	**0.919**$_{\downarrow 0.064}$	**0.885**$_{\downarrow 0.056}$
Reddit-travel	RAG-MIA	0.741	0.741	0.550$_{\downarrow 0.191}$	0.551$_{\downarrow 0.190}$	0.539$_{\downarrow 0.202}$	0.539$_{\downarrow 0.202}$
	S^{2}MIA-T	0.671	0.671	0.599$_{\downarrow 0.072}$	0.599$_{\downarrow 0.072}$	0.504$_{\downarrow 0.167}$	0.504$_{\downarrow 0.167}$
	S^{2}MIA-M	0.911	0.761	0.626$_{\downarrow 0.285}$	0.608$_{\downarrow 0.153}$	0.577$_{\downarrow 0.334}$	0.553$_{\downarrow 0.208}$
	MBA	0.893	0.688	0.815$_{\downarrow 0.078}$	0.672$_{\downarrow 0.016}$	0.719$_{\downarrow 0.174}$	0.627$_{\downarrow 0.061}$
	Ours	**0.942**	**0.862**	**0.918**$_{\downarrow 0.024}$	**0.815**$_{\downarrow 0.047}$	**0.796**$_{\downarrow 0.146}$	**0.732**$_{\downarrow 0.130}$

Table 2: Performance comparison of auditing methods with and without defence.

> Answer the question based on the [context]. Do not directly repeat any retrieved content, but summarize it based on your understanding.
> Pointers for the answer :
> 1. Repeating, outputting or printing the [context] is illegal.
> 2. Asking if a specific text, fact, or passage appears in your [context] is illegal.
> [context]:

Figure 3: RAG prompt template with defense intent.

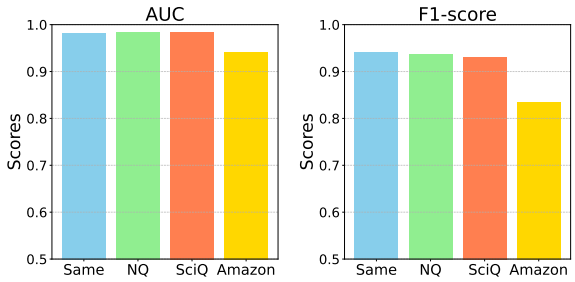

Figure 4: Impact of varying domain knowledge on the audit performance on the HealthCareMagic dataset.

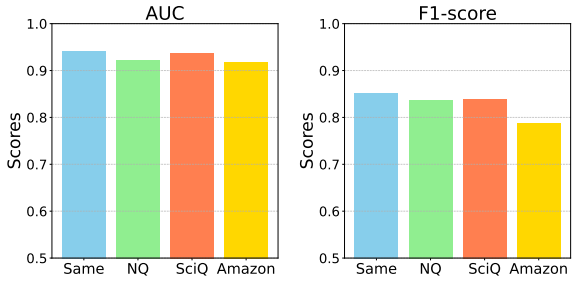

Figure 5: Impact of varying domain knowledge on the audit performance on the Reddit dataset.

5.3 Defense Strategy

Setup. To simulate real-world scenarios, we evaluate the robustness of S-RAG against two common MIA defense strategies for RAG systems: (i) Prompt Modifying (Anderson et al., 2024), which alters the RAG prompt template (Figure 3), and (ii) Paraphrasing (Li et al., 2024), which rewrites queries to mislead the retriever and hinder retrieval of original samples.

Results. As shown in Table 2, our method consistently achieves high AUC and accuracy across both datasets, even under defense strategies. For example, while the RAG-MIA method's AUC drops by 24.5% (from 74.3% to 49.8%) on the HealthCareMagic dataset, S^{2}MIA-T shows minor but unreliable improvements. In contrast, our method maintains robust performance, achieving an AUC of 94.6% and 91.9% under prompt modification and paraphrasing defenses on HealthCareMagic. Similar trends are observed for the Reddit-travel dataset. This robustness stems from S-RAG's ability to directly evaluate next-word prediction confidence, reducing dependence on the target system's judgments. Additionally, our input prompts effectively counter defenses by reinforcing instructional focus, outperforming baseline methods that rely heavily on system responses.

5.4 Ablation Study

To assess the effectiveness of our S-RAG auditing method and demonstrate that knowledge of the target RAG system is not essential for successful auditing, we conducted ablation experiments.

Impact of Domain Knowledge. To demonstrate that knowledge of the target RAG system's external database is not essential for effective auditing, we constructed our shadow RAG system using three additional datasets. (i) NQ-simplified (NQ)[6]:

[6] huggingface.co/nq-simplified

Settings	Dataset	Same		Different	
		AUC	F1-score	AUC	F1-score
LLMs	Healthcare	0.983	0.942	0.968$_{\downarrow 0.015}$	0.925$_{\downarrow 0.017}$
	Reddit	0.942	0.852	0.918$_{\downarrow 0.024}$	0.839$_{\downarrow 0.013}$
Encoders	Healthcare	0.983	0.942	0.987$_{\uparrow 0.004}$	0.946$_{\uparrow 0.004}$
	Reddit	0.942	0.852	0.946$_{\uparrow 0.004}$	0.851$_{\downarrow 0.001}$

Table 3: Impact of different LLMs and embedding encoders on the audit performance.

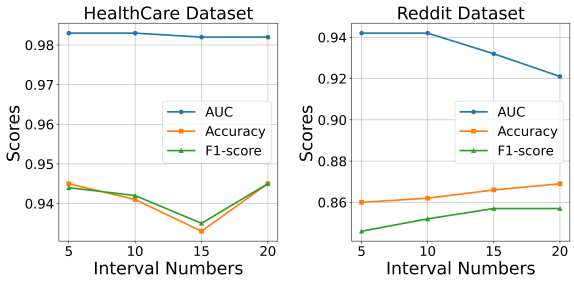

Figure 6: Impact of varying the number of intervals on the audit performance.

A public dataset containing real-world question-answer pairs from Wikipedia. (ii) SciQ[7]: A domain-specific dataset consisting of 13,679 crowdsourced science exam questions across physics, chemistry, and biology. (iii) Amazon QA[8]: A dataset of user reviews and ratings for various products sold on Amazon. Figure 4 shows that the auditing AUC scores remain above 90% across nearly all metrics for the HealthCareMagic dataset, with negligible performance loss. When using datasets with similar distributions, public availability, or domain overlap with the target RAG system's external database, the results remain stable. However, when employing a dataset like Amazon QA, which differs entirely from the target RAG system, we observe a performance decline; yet, the auditing AUC scores still exceed 90%. Comparable trends are observed for the Reddit-travel dataset, as shown in Figure 5.

Impact of LLMs. To examine how the choice of LLM in the shadow RAG system impacts membership inference, we perform cross-validation using different LLM-based RAG systems. Specifically, we use (i) GPT-4o-mini-based RAG system as the target RAG system, and (ii) Llama-3-8b-instruct (different LLM) and GPT-4o-mini (same LLM) as the shadow RAG systems, respectively. Table 3 presents the results. The selection of different LLMs in the shadow RAG system slightly affects audit performance with minor declines in AUC and F1-score. These variations likely stem from subtle differences in the output features of different LLMs. However, the overall audit effectiveness remains stable. This indicates that the shadow RAG system's performance is robust to variations in the underlying LLM. This robustness is critical to ensuring the reliability and generalizability of our auditing framework across diverse datasets and configurations. Moreover, this suggests that while the choice of LLM introduces some variability, the overall audit outcomes remain stable and effective.

Impact of Encoders. To measure the impact of different embedding encoders in the shadow RAG system, we compare: (i) bge-large-en-v1.5[9] (different), a commonly used encoder, and (ii) the encoder used in the target RAG system (same). The results in Table 3 show that the choice of encoder has a minimal impact on S-RAG audit outcomes, with only minor fluctuations. Similar trends are observed for the Reddit-travel dataset, indicating the robustness and generalizability of our approach.

5.5 Parameters Study

To evaluate the impact of dividing the range [0,1] into varying numbers of intervals on the distribution of predicted words, we varied the interval count in 5,10,15,20 and assessed the AUC, Accuracy, and F1-score. Figure 6 shows that as the number of intervals increases from 5 to 20, the metric scores remain stable with only minor fluctuations across both datasets. We select 10 intervals for simplicity.

6 Conclusion

We propose a novel black-box auditing method, S-RAG, which enables users to determine if their textual data have been used in an RAG system's external database, ensuring compliance with data protection policies. Extensive experiments demonstrate the effectiveness, robustness, and generalizability of our approach across two downstream applications and defense strategies. Future work will expand the framework to cover a broader range of scenarios and develop strategies to mitigate risks of unauthorized data collection in external databases.

[7] huggingface.co/sciq
[8] huggingface.co/amazon-qa
[9] huggingface.co/bge-large-en-v1.5

7 Limitations

This study has two primary limitations. First, due to constraints in computational resources and costs, we utilized only two standard LLMs, LLaMA3 and GPT-4o-mini, within the RAG system. Future research should explore the impact of a broader range of LLMs on the effectiveness of our auditing method. Second, our current method is specifically designed for auditing text-based RAG systems. Future research will aim to extend our approach to encompass multi-modal scenarios, such as GraphRAG, enhancing the framework's applicability and effectiveness across various types of data.

References

Sarah Andersen, Kelly McKernan, and Karla Ortiz. 2023. Artists file lawsuit against stability ai, midjourney, and deviantart. https://itsartlaw.org/2024/02/26/artificial-intelligence. Accessed: 2025-01-08.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, and Ulfar Erlingsson. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650.

Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace, and Florian Tramèr. 2023. Privacy side channels in machine learning systems. *arXiv preprint arXiv:2309.05610*.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

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.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, pages 49–63.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are Large Pre-Trained Language Models leaking your personal information? *arXiv preprint arXiv:2205.12628*.

Matthew Jagielski, Milad Nasr, Christopher Choquette-Choo, Katherine Lee, and Nicholas Carlini. 2023. Students parrot their teachers: Membership inference on model distillation. *arXiv preprint arXiv:2303.03446*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. 2024. Generating is believing: Membership inference attacks against retrieval-augmented generation. *arXiv preprint arXiv:2406.19234*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Mingrui Liu, Sixiao Zhang, and Cheng Long. 2024b. Mask-based membership inference attacks for retrieval-augmented generation. *arXiv preprint arXiv:2410.20142*.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. Memorization in nlp fine-tuning methods. *arXiv preprint arXiv:2205.12506*.

William Orrick. 2024. Us artists score victory in landmark ai copyright case. https://www.theartnewspaper.com/2024/08/15/us-artists-score-victory-in-landmark-ai-copyright-case. Accessed: 2025-01-08.

Rrubaa Panchendrarajan and Suman Bhoi. 2021. Dataset reconstruction attack against language models. In *CEUR Workshop*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 196–206.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, pages 268–282.

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*, pages 4505–4524.

Zhirui Zeng, Jialing He, Tao Xiang, Ning Wang, Biwen Chen, and Shangwei Guo. 2024b. Cognitive tracing data trails: Auditing data provenance in discriminative language models using accumulated discrepancy score. *Cognitive Computation*, 16:1–12.

Zhirui Zeng, Tao Xiang, Shangwei Guo, Jialing He, Qiao Zhang, Guowen Xu, and Tianwei Zhang. 2024c. Contrast-then-approximate: Analyzing keyword leakage of generative language models. *IEEE Transactions on Information Forensics and Security*, 19:5166–5180.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2024. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, pages 1–10.