WHEN SIMILARITY MISLEADS: QUERY PERTURBATION FOR ROBUST MEMBERSHIP INFERENCE IN RAG

Anonymous authors

ABSTRACT

Retrieval-Augmented Generation (RAG) systems enhance LLMs with external knowledge bases, but introduce new privacy risks. Membership inference attacks (MIAs) against RAG systems can exploit black-box access to infer whether a target document is contained in the knowledge base. Existing works typically rely on response accuracy, yet semantically similar non-member documents can also provide sufficient evidence for correct answers, causing false-positive inferences. To address this challenge, we propose QP-MIA, a black-box MIA based on query composition perturbation. QP-MIA constructs joint queries from target-specific *fill-in-the-blank* questions and perturbs their composition by replacing a subset of questions and reordering their positions, while preserving their connection to the target document. It then contrasts responses to two partially overlapping query compositions and extracts membership signals from changes in answer accuracy. This design distinguishes coherent evidence from genuine members from fragmented evidence provided by semantically similar non-members. Extensive experiments demonstrate an average AUC improvement of 20% over representative baselines across diverse RAG configurations.

1 INTRODUCTION

LLMs have demonstrated remarkable capabilities in knowledge understanding and generation, enabling a broad range of knowledge-intensive applications, such as question answering (Raffel et al., 2020), knowledge retrieval (Long et al., 2024), and intelligent customer service (Fuchs et al., 2026). Retrieval-Augmented Generation (RAG) further enhances these capabilities by coupling LLMs with external knowledge bases that may contain sensitive and proprietary information, such as medical records (Liang et al., 2025), financial documents (Iaroshev et al., 2024), and confidential enterprise data (Byun et al., 2025). By grounding generation in dynamic and domain-specific knowledge, RAG alleviates the limitations of stale parametric knowledge and insufficient domain coverage.

However, integrating external knowledge also introduces a new privacy attack surface. Even when the underlying knowledge base is inaccessible, its contents can influence the observable behavior of a RAG system. An adversary can exploit this behavior by issuing carefully crafted queries to a black-box RAG service and inferring whether a specific target document is included in its knowledge base. Such attacks, known as membership inference attacks (MIAs) against RAG (Wang et al., 2025), expose document-level membership information without requiring direct access to the underlying knowledge base, posing a fundamental privacy threat to RAG applications.

Existing MIAs against RAG systems (Liu et al., 2025; Guan et al., 2026) predominantly rely on answer correctness as the membership signal. An attacker constructs target-specific questions from a target document and queries the RAG system, based on the intuition that a member document is more likely to be retrieved and provide supporting evidence, resulting in higher answer accuracy. However, this intuition can break down when the knowledge base contains semantically similar non-member documents. Such documents may provide overlapping or fragmented evidence that is sufficient to answer individual target-specific questions correctly, even when the target document is absent. Consequently, high answer accuracy alone is not a reliable indicator of document membership, as it can also arise from semantically related evidence and lead to false-positive inferences.

A natural way to disentangle these factors is to perturb the query and examine how the RAG system responds. DCMI (Gao et al., 2025) explores this direction by perturbing individual target-specific

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

questions through token replacement and using the resulting response changes as membership signals, particularly to mitigate interference from semantically similar non-member documents. However, token-level perturbation faces an inherent trade-off. Weak perturbations may leave the retrieved evidence largely unchanged, providing limited additional membership information, whereas aggressive perturbations may substantially alter the query semantics, making the resulting response changes difficult to attribute specifically to document membership. This limits its ability to systematically alter the retrieved evidence while preserving the semantic connection to the target document.

Our key insight is to establish a different principle for probing membership. Instead of perturbing the semantics of an individual question, we perturb the semantic composition of the query while preserving its connection to the target document. Concretely, we formulate multiple target-specific questions into a joint query and construct controlled variants by replacing a subset of questions with alternative target-related questions and varying their ordering. Although the resulting queries differ in composition, they remain anchored to the same underlying semantics of the target document. This controlled change in query context exposes a fundamental difference between target-grounded and similarity-induced evidence. When the target document is a genuine member, its coherent evidence can support different query compositions, and the resulting answer accuracy remains relatively robust. When the target is a non-member, however, correct responses may depend on fragmented evidence distributed across semantically related documents; changing the query composition can alter the retrieved evidence and disrupt this fragmented support, leading to a larger drop in answer accuracy. Thus, rather than treating membership as a property reflected solely in the absolute correctness of responses, we probe membership through how answer accuracy responds to controlled changes in query composition.

Based on this insight, we propose QP-MIA, a black-box membership inference attack against RAG systems that probes document membership through response changes induced by controlled perturbations of query composition, thereby mitigating the interference from semantically similar non-member documents. QP-MIA consists of two key components, *Query Composition Perturbation* and *Contrastive Membership Signal Extraction*. The former constructs two partially overlapping joint queries composed of target-specific *fill-in-the-blank* questions, varying their question content and ordering while retaining shared questions as controlled anchors. The latter derives a membership signal by contrasting the responses to the two query compositions, jointly capturing answer accuracy and the changes induced by query perturbation. By exploiting response changes under controlled query composition changes, QP-MIA distinguishes coherent evidence grounded in a genuine member document from fragmented evidence arising from semantically similar non-member documents. Extensive experiments across diverse datasets and RAG configurations show that QP-MIA substantially improves membership inference performance, achieving an average improvement of 20% over representative baselines.

2 PRELIMINARIES

We firstly introduce RAG systems, followed by the problem definition and a review of related work.

2.1 RETRIEVAL-AUGMENTED GENERATION

RAG systems augment LLMs with an external knowledge base $\mathcal{D}$ that is queried at inference time. Given a user query q, the system retrieves relevant documents from $\mathcal{D}$ and incorporates them into the generation context to improve the accuracy and factuality of model outputs. A canonical RAG pipeline consists of two modules: (i) *Retriever.* The retriever $\mathcal{R}$ encodes documents in $\mathcal{D}$ into dense representations and, given q, retrieves the top-k documents with the highest semantic similarity: $D_k = \underset{d_i \in \mathcal{D}}{\arg \mathrm{TopK}}\ \mathrm{Sim}(q, d_i)$. (ii) *Generator*. The documents in D_k are combined with q using an instruction template and provided to the LLM, which generates the response: $\Omega = \mathrm{LLM}\big(\mathrm{Ins}(q, D_k)\big)$, where $\mathrm{Ins}(\cdot)$ organizes the query and retrieved documents into the generation prompt.

2.2 PROBLEM DEFINITION

Given a RAG system equipped with an external knowledge base $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, the adversary aims to determine whether a target document d^* is contained in $\mathcal{D}$. To this end, the adversary

constructs a set of target-specific queries based on d^* and submits them to the black-box RAG system. Given the resulting responses Ω, the adversary applies an inference algorithm $\mathcal{M}$ to determine the membership status of d^*. Formally, $\mathcal{M}(d^*, \Omega) \in \{0, 1\}$, where $\mathcal{M}(d^*, \Omega) = 1$ indicates that $d^* \in \mathcal{D}$ and $\mathcal{M}(d^*, \Omega) = 0$ otherwise.

Adversary's Capabilities. In black-box setting, the adversary can only interact with the target RAG system via external queries and observe the generated responses, without any access to the private knowledge base $\mathcal{D}$, the retrieval pipeline, or the internal parameters of the underlying LLM. Given a sequence of queries $(q_1, q_2, \ldots, q_m)$ specifically tailored to the target sample d^*, the adversary can observe the corresponding responses $\Omega = (a_1, a_2, \ldots, a_m)$ from the RAG system, where each a_i denotes the output generated in response to query q_i.

2.3 Existing works about MIAs against RAG

MIAs are a canonical privacy threat in machine learning (Hu et al., 2022), where an adversary infers whether a target sample was used to train a model from its observable outputs or parameters. RAG systems, however, introduce an additional membership leakage surface: their responses are generated not only from the LLM's parametric knowledge but also from an external knowledge base accessed at inference time. Thus, beyond inferring training-set membership of the underlying LLM (Meeus et al., 2024), an adversary may infer whether a target document is contained in the RAG knowledge base, posing a distinct privacy threat. With the growing adoption of RAG, recent studies have begun to investigate such document-level membership inference attacks (Feng et al., 2025). From the perspective of the membership signal they exploit, existing document-level MIAs against RAG systems can be organized into three categories: (i) target-specific correctness probing, (ii) response-content and behavioral analysis, and (iii) retrieval-interference-aware calibration.

Target-specific correctness probing. Methods in this category actively construct target-specific probing tasks and treat the RAG system's ability to correctly recover target-specific information as membership evidence. RAG-MIA (Anderson et al., 2025) inserts the target document into a verification prompt and interprets affirmative responses as membership indicators. IA (Naseh et al., 2025) and E-MIA (Guan et al., 2026) use an auxiliary LLM to generate target-specific questions and infer membership from answer correctness, with IA relying on true-or-false questions and E-MIA on structured exams. MBA (Liu et al., 2025) masks key tokens in the target document and uses masked-token recovery accuracy as the membership score. MEntA (Nguyen et al., 2026) further adopts natural-language queries to enable low-cost probing with only a small number of interactions.

Response-content and behavioral analysis. A second line of work analyzes properties of the generated responses rather than discrete correctness on target-specific probes. S^2MIA (Li et al., 2025b) distinguishes members from non-members through differences in semantic similarity and perplexity between the target document and the system output. BudgetLeak (Li et al., 2025a) instead examines how response quality evolves under different generation budgets, characterizing membership via a behavioral side-channel. Although these approaches capture finer-grained correlations between the target and system behavior, neither they nor correctness-based methods explicitly address interference from similar non-member documents, which can still yield target-like responses and unstable membership decisions.

Retrieval-interference-aware calibration. DCMI (Gao et al., 2025) is representative of methods that explicitly target retrieval interference from similar non-member documents. It applies token-level perturbations to individual target-specific questions and treats the induced response changes as membership evidence, aiming to reduce false positives caused by partially relevant neighbors. However, token-level edits are hard to balance: mild changes often leave retrieval largely unaffected, whereas stronger edits risk shifting query semantics and confounding the membership signal. In addition, the method still relies on confidence-based scores available only under gray-box access, and is therefore largely ineffective in a pure black-box setting.

In summary, existing MIA methods for RAG systems mainly focus on target-specific probing, response-based membership signals, and retrieval-interference-aware calibration. However, each line of work still has practical limitations. Motivated by these observations, we develop a black-box MIA method based on query composition perturbation that reduces the impact of similar non-member documents.

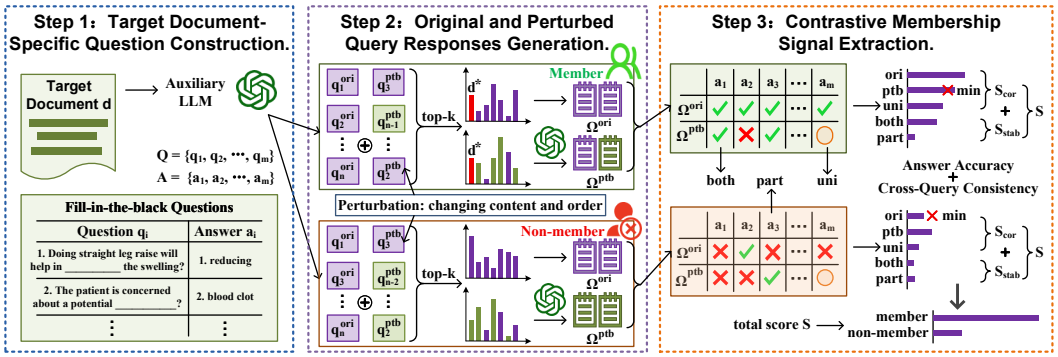


Figure 1: The overall architecture of QP-MIA.

3 QP-MIA

In this section, we introduce QP-MIA, a black-box membership inference attack against RAG systems based on query composition perturbation. Rather than relying on the absolute correctness or similarity of individual responses, QP-MIA probes membership by examining response changes induced by controlled perturbations of target-specific query composition. In the following, we first provide an overview of the QP-MIA framework, followed by a detailed description of its key steps.

3.1 Overview

As shown in Fig. 1, given a target document d^*, QP-MIA first generates a set of target-specific questions along with their corresponding answers. Based on these question-answer pairs, it then constructs a joint query by combining multiple questions and generates a perturbed query by modifying their composition. The two queries are submitted to the target RAG system to obtain the corresponding responses. Finally, QP-MIA performs *Contrastive Membership Signal Extraction* by comparing the responses to derive the membership signal. The details of each component are described as follows.

Step 1: Target Document-Specific Question Construction. Given a target document d^*, QP-MIA probes its membership by comparing responses under controlled perturbations of query composition. This requires a set of target-specific questions that can be structurally manipulated while preserving their semantic connection to d^*. We therefore adopt *fill-in-the-blank* questions rather than binary yes/no questions. By encoding explicit and fine-grained information units, fill-in-the-blank questions can be selectively replaced or reordered while maintaining the underlying information-seeking intent, making them suitable for constructing controlled variations of the query composition.

Specifically, QP-MIA employs an auxiliary LLM to transform the target document d^* into a set of document-specific fill-in-the-blank questions $\mathcal{Q} = \{q_1, q_2, \ldots, q_m\}$ and their corresponding ground-truth answers $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$. Each question-answer pair (q_i, a_i) is derived from information contained in d^*, enabling the correctness of each RAG response to be directly evaluated against its corresponding ground-truth answer. To ensure the quality and relevance of the generated questions, we prompt the auxiliary LLM to act as a professional exam designer, with the complete prompt provided in App. A.1.

Step 2: Original and Perturbed Query Responses Generation. Given the generated question set $\mathcal{Q}$, QP-MIA constructs a joint query $\mathcal{Q}^{\text{ori}}$ by combining several questions from $\mathcal{Q}$ and generates its perturbed counterpart $\mathcal{Q}^{\text{ptb}}$ using the proposed query composition perturbation technique. The two queries are submitted separately to the target RAG system, which is instructed to answer each constituent question individually, yielding the original response set Ω^{ori} and the perturbed response set Ω^{ptb}. The step will be detailed later.

Step 3: Contrastive Membership Signal Extraction. Given Ω^{ori} and Ω^{ptb}, QP-MIA jointly analyzes the answer accuracy of the two response sets and the response changes on shared questions

across the two query compositions to derive a membership signal for the target document d^*. The resulting signal is compared with a threshold τ to determine whether d^* is contained in the knowledge base. The step will be detailed later.

3.2 Original and Perturbed Query Responses Generation

In this step, given the generated question set $\mathcal{Q}$, QP-MIA constructs a joint query by combining multiple target-specific questions. This joint composition serves as a structured basis for generating a controlled perturbed counterpart, allowing the retrieval context to be deliberately varied while preserving the underlying connection to the target document. The resulting pair of query compositions provides two complementary views of the same target, whose responses can subsequently be exploited to derive a stronger membership signal.

Specifically, given $\mathcal{Q} = \{q_1, q_2, \ldots, q_m\}$, QP-MIA first randomly selects n target-specific questions to form a subset $\mathcal{Q}^{\mathrm{ori}} = \{q_1^{\mathrm{ori}}, \ldots, q_n^{\mathrm{ori}}\} \subset \mathcal{Q}$. These questions are then concatenated to construct the original joint query:

$$q^{\mathrm{ori}} = \mathrm{Concat}\left(q_1^{\mathrm{ori}}, \ldots, q_n^{\mathrm{ori}}\right).$$

The resulting joint query serves as the reference composition, where multiple target-specific questions are presented together and jointly determine the retrieval context.

Query Composition Perturbation. We then construct a perturbed counterpart of q^{ori} by modifying its query composition. The goal is to alter the retrieval context while preserving the underlying semantic connection to the target document. To this end, as illustrated in the second step of Fig. 1, we introduce perturbations along two dimensions: *question content* and *question order*.

Specifically, we first retain a subset of questions from $\mathcal{Q}^{\mathrm{ori}}$ as shared questions, denoted by $\mathcal{Q}^{\mathrm{ov}}$. These shared questions serve as controlled anchors, ensuring that the original and perturbed query compositions maintain a common semantic basis. We then randomly sample $|\mathcal{Q}^{\mathrm{ori}} \setminus \mathcal{Q}^{\mathrm{ov}}|$ questions from the remaining question pool $\mathcal{Q} \setminus \mathcal{Q}^{\mathrm{ori}}$ and use them to replace the non-shared questions in $\mathcal{Q}^{\mathrm{ori}} \setminus \mathcal{Q}^{\mathrm{ov}}$. This yields the perturbed question set $\mathcal{Q}^{\mathrm{ptb}}$, satisfying $\mathcal{Q}^{\mathrm{ori}} \cap \mathcal{Q}^{\mathrm{ptb}} = \mathcal{Q}^{\mathrm{ov}}$. This process introduces a controlled perturbation to the question content while preserving a subset of common target-specific questions between the two query compositions.

We further perturb the question order by applying a permutation π to the questions in $\mathcal{Q}^{\mathrm{ptb}}$, yielding the perturbed joint query

$$q^{\mathrm{ptb}} = \mathrm{Concat}\left(q_{\pi(1)}^{\mathrm{ptb}}, \ldots, q_{\pi(n)}^{\mathrm{ptb}}\right).$$

Together, question replacement and reordering modify the composition and presentation of the joint query, thereby providing controlled changes to the retrieval context while preserving its semantic connection to the target document. Such perturbations can expose membership-dependent retrieval behavior, as illustrated by the member and non-member paths in Fig. 1: a genuine member document can provide coherent evidence for multiple target-specific questions across different query compositions, whereas a similarity-confounded non-member may rely on evidence distributed across semantically related documents, making its responses more sensitive to changes in query composition.

Finally, the original and perturbed joint queries are separately submitted to the target RAG system. For each query, the system retrieves its corresponding top-k documents and is instructed to answer each constituent question individually. Let $D_k(q^{\mathrm{ori}})$ and $D_k(q^{\mathrm{ptb}})$ denote the retrieved document sets for the original and perturbed queries, respectively. The corresponding response sets are obtained as

$$\Omega^{\mathrm{ori}} = \mathrm{LLM}\left(\mathrm{Ins}\left(q^{\mathrm{ori}}, D_k(q^{\mathrm{ori}})\right)\right), \qquad \Omega^{\mathrm{ptb}} = \mathrm{LLM}\left(\mathrm{Ins}\left(q^{\mathrm{ptb}}, D_k(q^{\mathrm{ptb}})\right)\right).$$

Thus, QP-MIA obtains two response sets under different query compositions. By comparing Ω^{ori} and Ω^{ptb}, the subsequent *Contrastive Membership Signal Extraction* module captures the response

changes induced by query composition perturbation and derives the membership signal for the target document d^*.

3.3 CONTRASTIVE MEMBERSHIP SIGNAL EXTRACTION

Given the responses to the original and perturbed joint queries, $\Omega^{\mathrm{ori}} = \{\Omega_1^{\mathrm{ori}}, \ldots, \Omega_n^{\mathrm{ori}}\}$ and $\Omega^{\mathrm{ptb}} = \{\Omega_1^{\mathrm{ptb}}, \ldots, \Omega_n^{\mathrm{ptb}}\}$, let Ω_i^{ori} and Ω_i^{ptb} denote the answers to the i-th question under the original and perturbed query compositions, respectively. QP-MIA derives the membership signal by contrasting the two response sets, jointly considering overall answer accuracy and the consistency of answers to shared questions across the two query compositions, rather than relying solely on absolute answer accuracy.

The underlying principle is that document membership is reflected in how retrieved evidence supports target-specific questions under different query compositions. For a genuine member d^*, the retrieved document sets $D_k(q^{\mathrm{ori}})$ and $D_k(q^{\mathrm{ptb}})$ are more likely to contain d^*, allowing the target document to provide coherent evidence for multiple target-specific questions. In contrast, for a similarity-confounded non-member, semantically related documents may still provide sufficient evidence for individual questions, but such evidence is more likely to be fragmented across different documents and cover only subsets of the target-specific information. Consequently, changing the query composition can alter the retrieved evidence and induce corresponding changes in answer accuracy. By jointly capturing overall answer accuracy and the behavior of shared questions across the two compositions, QP-MIA extracts a membership signal that distinguishes coherent evidence grounded in the target document from fragmented evidence arising from semantically similar non-member documents. To this end, as illustrated in the third step of Fig. 1, QP-MIA analyzes two complementary signals: *Answer Accuracy* and *Cross-Query Consistency*, as detailed below.

Answer Accuracy. Let C^{ori} and C^{ptb} denote the numbers of correctly answered questions in $\mathcal{Q}^{\mathrm{ori}}$ and $\mathcal{Q}^{\mathrm{ptb}}$, respectively. Let C^{uni} denote the number of correctly answered questions that appear in only one of the two query compositions, i.e., questions in $\left(\mathcal{Q}^{\mathrm{ori}} \cup \mathcal{Q}^{\mathrm{ptb}}\right) \setminus \mathcal{Q}^{\mathrm{ov}}$. We determine whether a question is correctly answered based on the token-level overlap between the generated response Ω_i and its reference answer a_i. Specifically, we compute $o(\Omega_i, a_i) = \frac{|T(\Omega_i) \cap T(a_i)|}{\max(|T(\Omega_i)|, |T(a_i)|)}$, where $T(\cdot)$ denotes the set of tokens in an answer. A question is considered correctly answered if $o(\Omega_i, a_i) \geq \theta_{\mathrm{cor}}$, where θ_{cor} is a predefined correctness threshold.

Based on these binary correctness judgments, we define the correctness score by combining the response accuracy under the original and perturbed query compositions with a positive weight λ_1:

$$S_{\mathrm{cor}} = \lambda_1 \left(\max\left(C^{\mathrm{ori}}, C^{\mathrm{ptb}}\right) + C^{\mathrm{uni}} \right).$$

The term $\max(C^{\mathrm{ori}}, C^{\mathrm{ptb}})$ captures the highest answer accuracy achieved under the two query compositions, while C^{uni} accounts for correctly answered questions introduced through query perturbation. A higher S_{cor} indicates that the retrieved evidence under at least one query composition provides broader support for the target-specific information, which is more likely when the target document itself is retrieved.

Cross-Query Consistency. For the shared questions $\mathcal{Q}^{\mathrm{ov}}$, which appear in both query compositions, we further examine whether the same target-specific facts can be correctly recovered under both compositions. Let C^{both} denote the number of shared questions that are answered correctly under *both* compositions, and let C^{part} denote the number of shared questions that are answered correctly under *exactly one* of the two compositions. With a positive weight $\lambda_2 > 1$, we define

$$S_{\mathrm{stab}} = \lambda_2 C^{\mathrm{both}} + C^{\mathrm{part}}.$$

The two terms capture different degrees of cross-query consistency. C^{both} represents target-specific facts that can be correctly recovered under both query compositions, indicating that the underlying evidence remains effective despite changes in query composition. In contrast, C^{part} captures partial recovery, where a shared question is answered correctly under only one composition. We assign

Table 1: AUC of different methods under varied generators, retrievers, and RAG frameworks.

Setting		TREC-COVID					HealthCareMagic					MS-MARCO				
		RAG	E-I	E-J	DCMI	QP	RAG	E-I	E-J	DCMI	QP	RAG	E-I	E-J	DCMI	QP
Generator	Llama	0.600	0.741	0.639	0.690	**0.935**	0.542	0.605	0.731	0.753	**0.944**	0.529	0.696	0.591	0.573	**0.806**
	Mistral	0.790	0.765	0.818	0.810	**0.928**	0.730	0.640	0.894	0.775	**0.955**	0.510	0.684	0.677	0.677	**0.804**
	Qwen	0.799	0.807	0.882	0.805	**0.942**	0.725	0.571	0.813	0.713	**0.881**	0.625	0.677	0.734	0.623	**0.837**
Retriever	Contriever	0.540	0.756	0.622	0.660	**0.914**	0.540	0.618	0.748	0.753	**0.870**	0.530	0.620	0.607	0.625	**0.785**
Framework	SC-RAG	0.565	0.689	0.570	0.610	**0.843**	0.538	0.590	0.667	0.680	**0.822**	0.511	0.607	0.565	0.500	**0.818**
	Spring	0.500	0.744	0.597	0.645	**0.906**	0.500	0.602	0.745	0.635	**0.867**	0.523	0.603	0.576	0.495	**0.771**

a larger weight to C^{both} to emphasize correctness that persists across the controlled perturbation, while retaining C^{part} as complementary membership evidence.

Accordingly, the final membership score is defined as $S = S_{\mathrm{cor}} + S_{\mathrm{stab}}$. A larger S indicates stronger membership evidence. Given a decision threshold τ, QP-MIA predicts d^* as a member if $S > \tau$, and as a non-member otherwise.

4 Evaluation

Setup. We evaluate QP-MIA, denoted as QP, on three benchmark datasets, TREC-COVID (Thakur et al., 2021), HealthCareMagic (Wang), and MS-MARCO (Nguyen et al., 2016), which serve as the knowledge bases for the evaluated RAG systems. We compare QP with three representative RAG membership inference attacks: RAG-MIA (Anderson et al., 2025), denoted as RAG, DCMI (Gao et al., 2025), and E-MIA (Guan et al., 2026). For E-MIA, we consider two variants, E-MIA-I (E-I) and E-MIA-J (E-J), which issue multiple target-specific questions individually and jointly, respectively. Unless otherwise specified, we use Llama-2-7B-chat (Touvron et al., 2023) as the generator, E5-base as the retriever, Standard RAG (Ram et al., 2023) with top-k=4, and report AUC (Sankararaman et al., 2009) as the primary metric together with ACC (Choquette-Choo et al., 2021) under a fixed threshold. For each target document, we generate $m = 7$ fill-in-the-blank questions and submit two composite queries of length $n = 5$ that share three questions and differ in the remaining four. Full dataset descriptions, baseline implementations, RAG configurations, metrics, and hyper-parameters are provided in App. B.

4.1 Main Results

We systematically evaluate QP-MIA (QP) under different RAG configurations. Specifically, we study the influence of generative modules, retrieval modules, RAG frameworks, and the number of retrieved documents respectively.

Impact of Generative Modules. To evaluate the robustness of QP-MIA across different generators, we fix the retriever and RAG framework to E5 and Standard RAG, respectively, and vary the generator among Llama, Mistral, and Qwen. As shown in Table 1, QP-MIA achieves the highest AUC across all three datasets and all evaluated generators. For example, with Llama on TREC-COVID, QP-MIA improves the AUC by approximately 35% over DCMI and 25% over E-MIA. The consistent gains over RAG-MIA and E-MIA stem from QP-MIA's complementary membership signals, which jointly exploit answer correctness and cross-query consistency under controlled perturbations of query content and order, rather than relying solely on answer accuracy under a fixed query composition. Compared with DCMI, which explicitly addresses interference from semantically similar non-member documents, QP-MIA adopts a query composition perturbation strategy that varies the retrieval context while preserving the underlying semantic connection to the target document. In contrast, DCMI relies on word-level perturbations of individual queries, making its effectiveness sensitive to the perturbation magnitude. Moreover, DCMI primarily leverages changes in response confidence as a membership signal. Although DCMI can operate in black-box settings by analyzing observable response changes, its performance degrades substantially when only text responses are available. QP-MIA instead derives membership signals entirely from text responses, enabling effective inference under strict black-box access. Similar trends are observed with Mistral and Qwen, where QP-MIA consistently achieves the highest AUC across all three datasets.

Table 2: ACC of different methods under varied generators.

Generator	TREC-COVID					HealthCareMagic					MS-MARCO				
	RAG	E-I	E-J	DCMI	QP	RAG	E-I	E-J	DCMI	QP	RAG	E-I	E-J	DCMI	QP
Llama	0.600	0.565	0.505	0.690	**0.893**	0.535	0.510	0.505	0.750	**0.893**	0.530	0.565	0.520	0.570	**0.680**
Mistral	0.790	0.585	0.650	0.810	**0.887**	0.730	0.520	0.715	0.775	**0.927**	0.510	0.600	0.575	0.675	**0.680**
Qwen	0.795	0.655	0.820	0.805	**0.873**	0.725	0.515	0.750	0.705	**0.847**	0.625	0.600	0.655	0.620	**0.693**

Besides, Table 2 reports the ACC of all methods across different generators. QP-MIA consistently achieves the highest ACC, outperforming the strongest baseline by approximately 20%–30% across datasets under Llama, with positive gains also observed under Mistral and Qwen. These results are consistent with the AUC results in Table 1, further demonstrating the effectiveness of QP-MIA across different generators.

Impact of Retrieval Modules. To evaluate the effectiveness of QP-MIA across different retrievers, we fix the generator and RAG framework to Llama and Standard RAG, respectively, and compare E5 with Contriever. As shown in Table 1, QP-MIA achieves the highest AUC across both retrievers and all three datasets. The first line in the table corresponds to E5, while the fourth line corresponds to Contriever. Under E5, QP-MIA outperforms the strongest baseline by approximately 20%–40% across datasets. Under Contriever, the performance gains over competing methods become smaller, yet QP-MIA still achieves the highest AUC, with improvements of approximately 20% over the E-MIA variants and 35% over DCMI on TREC-COVID. The higher absolute performance under E5 can be attributed to its stronger retrieval quality, as more accurate top-k retrieval makes target-specific evidence more likely to appear under both the original and perturbed query compositions. Nevertheless, QP-MIA maintains a clear advantage even under Contriever, demonstrating that its effectiveness is not tied to a specific retriever and remains robust across different retrieval settings.

Impact of RAG Frameworks. With Llama and Contriever as the generator and retriever, respectively, we evaluate QP-MIA and its competitors across three RAG frameworks: Standard RAG, SC-RAG, and Spring (fourth, fifth, and sixth lines in Table 1, respectively). Consistent with previous results, QP-MIA consistently achieves the highest AUC, improving over the strongest baseline by approximately 20%–25% on TREC-COVID and comparable margins on HealthCareMagic and MS-MARCO. While framework changes affect all methods, SC-RAG generally yields lower AUC, likely due to context compression that may discard target-relevant evidence, whereas Spring performs closer to Standard RAG. Nevertheless, QP-MIA maintains a clear advantage across all three frameworks, demonstrating its robustness to different RAG architectures.

Impact of Retrieved Document Numbers.

We evaluate QP-MIA with varying retrieval sizes $k \in [2, 6]$ on the HealthCareMagic dataset. As shown in Fig. 2, QP-MIA consistently achieves the highest AUC across all settings, maintaining an AUC of approximately 0.95 and outperforming the strongest baseline by more than 15%. As k increases, RAG-MIA exhibits a steady performance decline, while DCMI shows a milder downward trend, likely due to increasing interference from noisy or irrelevant retrieved content. The two E-MIA variants fluctuate without a consistent trend. In contrast, QP-MIA remains largely stable across different retrieval sizes and even slightly improves at larger k. These results demonstrate that QP-MIA remains effective and robust under varying retrieval sizes.

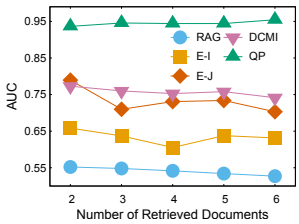
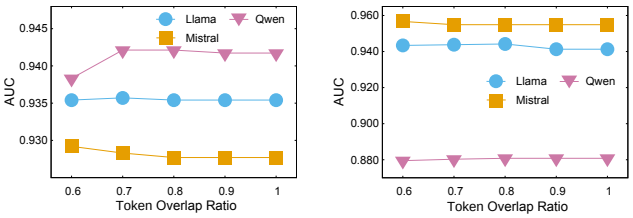


Figure 2: Performance on different numbers of retrieved documents.

4.2 Ablation Studies

Effect of correctness decision threshold. We vary the token-overlap threshold θ_{cor} used to determine whether a generated answer is considered correct, ranging from 0.60 to 1.00. As shown in Fig. 3, the AUC remains largely unchanged across this range on both

TREC-COVID and HealthCareMagic, consistently across all three generators. This robustness can be attributed to the short and fact-oriented nature of our fill-in-the-blank references, which primarily consist of concrete information such as numbers, names, and key phrases. Consequently, moderate variations in θ_{cor} rarely change the resulting correctness judgments or membership rankings. We therefore use $\theta_{\text{cor}} = 0.8$ as the default setting.



(a) TREC-COVID (b) HealthCareMagic

Figure 3: Effect of correctness decision threshold.

Comparison of different query compositions. We compare QP-MIA with two alternative query compositions in Table 3. *Single* constructs one joint query containing five randomly selected target-specific questions and uses its correctness as the membership signal. *Disjoint* performs two independent probes with five non-overlapping questions each and aggregates their correctness scores. The results show that incorporating overlapping questions in QP-MIA consistently improves AUC over both alternatives. For example, with Llama on HealthCareMagic, QP-MIA achieves an AUC of 0.94, compared with 0.89 and 0.88 for *Single* and *Disjoint*, respectively. Similar improvements are observed on TREC-COVID and MS-MARCO. These results demonstrate that contrasting responses to shared questions across different query compositions provides a more informative membership signal than aggregate correctness alone.

Table 3: AUC under different questioning strategies. Results are evaluated with E5 and Std. RAG.

Generator	TREC-COVID			HealthCareMagic			MS-MARCO		
	Single	Disjoint	QP	Single	Disjoint	QP	Single	Disjoint	QP
Llama	0.892	0.919	**0.935**	0.892	0.882	**0.944**	0.765	0.767	**0.806**
Mistral	0.895	0.904	**0.928**	0.902	0.894	**0.955**	0.746	0.753	**0.804**
Qwen	0.909	0.916	**0.942**	0.787	0.845	**0.881**	0.763	0.793	**0.837**

4.3 Evaluation against Potential Defenses

We further evaluate QP-MIA against three representative defenses: *Instr.*, which incorporates safety constraints into the system prompt; *Paraph.*, which rewrites submitted questions while preserving their semantic intent; and *Entity-Rel.*, which converts retrieved passages into structured triples before generation. We use *None* to denote the setting without any defense. Detailed defense configurations and complete analysis are provided in App. C.2.

Table 4: AUC before and after defense.

Defense	RAG	E-I	E-J	DCMI	QP
None	0.542	0.605	0.731	0.753	**0.944**
Instr.	0.533	0.628	0.699	0.707	**0.950**
Paraph.	0.527	0.594	0.708	0.560	**0.879**
Entity-Rel.	0.515	0.551	0.615	0.518	**0.792**

As shown in Table 4, Instr. has limited impact on the evaluated attacks, with their AUCs remaining largely unchanged. In contrast, Paraph. and Entity-Rel. lead to more pronounced performance degradation across most attack methods. Among them, DCMI experiences the largest drop and approaches random-guessing performance under the strongest defense. QP-MIA, in comparison, exhibits a more moderate degradation and maintains the highest AUC score across all evaluated defense settings.

5 Conclusion

We propose QP-MIA, a black-box membership inference attack against RAG systems that exploits response changes induced by controlled query-composition perturbations, in order to mitigate interference from semantically similar non-member documents. It consists of a carefully designed query perturbation strategy and a corresponding membership-signal extraction mechanism that jointly capture informative response changes for membership inference. Future work will extend our perturbation-based attack framework to multimodal RAG systems.

AI USE STATEMENT

In this work, we used generative AI tools only to assist with writing and translation, including polishing English phrasing and improving clarity of presentation. We have not used generative AI tools for research ideation, experimental design, data analysis, result interpretation, or the generation of scientific claims. We have carefully reviewed all AI-assisted text and take full responsibility for the final content of this paper, including all claims and conclusions.

ETHICS STATEMENT

The research in this paper focuses on MIA against RAG systems, aiming to evaluate potential privacy leakage risks in retrieval-augmented generation systems and promote the development of more secure and reliable LLM applications. All experiments in this paper are conducted based on publicly available datasets and models, without involving any real user data containing personally identifiable information. The proposed MIA method is only evaluated in controlled environments and does not access unauthorized private data, commercial systems, or real-world deployed environments.

REPRODUCIBILITY STATEMENT

To foster reproducibility, we provide details of our experimental setup in §4 and App. B, and include all prompts used in our experiments in App. A. We release our code and the datasets used in this work as a compressed archive in the supplementary material, together with configuration files for each experiment and a README with instructions on how to reproduce our results.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments.

REFERENCES

Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy, ICISSP*, 2025.

Grace Byun, Shinsun Lee, Nayoung Choi, and Jinho D. Choi. Secure multifaceted-rag for enterprise: Hybrid knowledge retrieval with security filtering. *CoRR*, abs/2504.13425, 2025.

Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

Kaiyue Feng, Guangsheng Zhang, Huan Tian, Heng Xu, Yanjun Zhang, Tianqing Zhu, Ming Ding, and Bo Liu. Ragleak: Membership inference attacks on rag-based large language models. In *Information Security and Privacy - 30th Australasian Conference, ACISP*, 2025.

Gilad Fuchs, Leonid Ekimov, Fei Dong, Jiahong Xie, Wei Liu, Maxim Manco, and Alexander Nus. Optimizing retrieval-augmented generation for e-commerce how-to assistance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL*, 2026.

Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. *DCMI:* A differential calibration membership inference attack against retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS*, 2025.

Zelin Guan, Shengda Zhuo, Zeyan Li, Jinchun He, Wangjie Qiu, Zhiming Zheng, and Shuqiang Huang. E-MIA: exam-style black-box membership inference attacks against RAG systems. *CoRR*, abs/2605.00955, 2026.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s):235:1–235:37, 2022.

Ivan Iaroshev, Ramalingam Pillai, Leandro Vaglietti, and Thomas Hanne. Evaluating retrieval-augmented generation models for financial report question and answering. *Applied Sciences*, 14 (20), 2024. ISSN 2076-3417.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

Hao Li, Jiajun He, Guangshuo Wang, Dengguo Feng, Zheng Li, and Min Zhang. Budgetleak: Membership inference attacks on RAG systems via the generation budget side channel. *CoRR*, abs/2511.12043, 2025a.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2023.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2025b.

Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. RGAR: recurrence generation-augmented retrieval for factual-aware medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Lan Liu, Yi Wang, Gaoyang Liu, Kai Peng, and Chen Wang. Membership inference attacks against machine learning models via prediction sensitivity. *IEEE Trans. Dependable Secur. Comput.*, 20 (3):2341–2347, 2023.

Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025, WWW*, 2025.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. Generative multi-modal knowledge retrieval with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, 2024.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium, USENIX Security*, 2024.

Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS*, 2025.

Nguyen Linh Bao Nguyen, Wanlun Ma, Viet Vo, Alsharif Abuadbba, Minghong Fang, Jun Zhang, and Yang Xiang. Five queries are enough: Query-efficient and surrogate-free membership inference attacks on RAG via entailment. *CoRR*, abs/2605.24312, 2026.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems, NIPS*, 2016.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331, 2023.

Sriram Sankararaman, Guillaume Obozinski, Michael I. Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967, 2009.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

Guangshuo Wang, Jiajun He, Hao Li, Min Zhang, and Dengguo Feng. Rag-leaks: difficulty-calibrated membership inference attacks on retrieval-augmented generation. *Sci. China Inf. Sci.*, 68(6), 2025.

Rongsheng Wang. Healthcaremagic-100k-en. Hugging Face. URL `https://huggingface.co/datasets/wangrongsheng/HealthCareMagic-100k-en`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

Yutao Zhu, Zhaoheng Huang, Zhicheng Dou, and Ji-Rong Wen. One token can help! learning scalable and pluggable virtual tokens for retrieval-augmented large language models. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, 2025.

A PROMPTS AND EXAMPLE TEXTS

In this section we list the prompts used throughout our experiments. App. A.1 presents the prompt for generating target-specific fill-in-the-blank exam questions, and App. A.2 presents the prompt used by the RAG system when answering the joint queries.

A.1 EXAM GENERATION PROMPT

Exam Generation Prompt

You are a professional exam designer. Generate EXACTLY 7 high-quality fill-in-the-blank questions based strictly on the provided document.
Document: {context}
Requirements:

- Generate EXACTLY 7 different questions.
- Each question must be a fill-in-the-blank sentence with a blank marked as _____.
- Do not use interrogative forms such as "What is ...?", "How long ...?", or "Which ...?".
- The blank should replace a specific short fact from the document (name, number, date, entity, term, etc.).
- Each answer must be a short substring from the document, preferably 1–5 words.
- Questions must require the document; do not ask common-knowledge facts alone.
- Do not invent information not in the document.
- Avoid duplicate or near-duplicate questions.
- Output ONLY valid JSON. No explanations, markdown, or extra text.

Good example:

```
{
  "question": "The company was founded in _____.",
  "answer": "1998"
}
```

Output format:

```
{
  "fill_in_blank": [
    {"question": "Sentence with _____.", "answer": "exact answer"},
    {"question": "Sentence with _____.", "answer": "exact answer"},
    ...
  ]
}
```

A.2 EXAM ANSWERING PROMPT

Exam Answering Prompt

Context: {context}
Please answer all fill-in-the-blank questions strictly:

- Output only the answer for each question, one per line.
- Use the format: 1. answer
- No explanations or extra text.
- If unsure, give the most likely short answer.

Questions: {questions_text}
Answers:
1.

B Additional Experimental Details

Datasets. We evaluate QP-MIA on three widely used open-domain datasets, TREC-COVID (Thakur et al., 2021), HealthCareMagic-100k-en (Wang), and MS-MARCO (Nguyen et al., 2016), which serve as the external knowledge bases of the RAG system. The three datasets cover scientific, biomedical, and general web domains. TREC-COVID and HealthCareMagic contain approximately 17K and 11K documents, respectively. MS-MARCO is a large-scale machine reading comprehension and question-answering corpus and the original release reports about 8.84 million passages. To better match realistic MIA targets, which typically contain a non-trivial amount of information, we remove documents with fewer than 50 tokens before building the knowledge bases.

Following prior RAG membership inference studies (Liu et al., 2023), we partition each dataset into member and non-member documents. Specifically, 80% of the documents are included in the RAG knowledge base as members, while the remaining 20% are held out as non-members and never indexed by the RAG system. For attack evaluation, we randomly sample 500 member and 500 non-member documents from each dataset, resulting in a balanced evaluation set of 1,000 target documents per dataset.

Baselines. We compare QP-MIA with a range of existing RAG membership inference attacks.

- ***RAG-MIA*** (Anderson et al., 2025) inserts the target document into a verification prompt and infers membership based on whether the RAG system confirms that the document appears in the retrieved context.
- ***DCMI*** (Gao et al., 2025) applies differential calibration by comparing the confidence of "Yes" responses under the original query and a semantically perturbed query, thereby reducing interference from similar non-member-retrieved documents.
- ***E-MIA*** (Guan et al., 2026) generates document-specific exams composed of objectively verifiable questions and determines membership according to the model's performance on these questions.

For a controlled comparison between individual and joint querying, we further implement two variants of E-MIA under the same exam construction: ***E-MIA-I***, which submits each question separately, and ***E-MIA-J***, which concatenates all questions into a single joint query.

Target RAG System. To evaluate QP-MIA, we establish a local RAG environment and evaluate our attack under three widely used RAG frameworks, including Standard RAG (Ram et al., 2023), SC-RAG (Li et al., 2023), and Spring (Zhu et al., 2025). We consider three representative generative modules: Llama-2-7B-chat (Touvron et al., 2023), Mistral-7B-Instruct (Jiang et al., 2023), and Qwen-2.5-7B (Yang et al., 2024), covering different open-source LLM architectures. For the retrieval component, we employ two widely used dense retrieval models, E5-base and Contriever, to construct the knowledge retrieval pipeline.

Metrics. We adopt the Area Under the ROC Curve (AUC) (Sankararaman et al., 2009) as the primary evaluation metric. AUC measures the probability that a randomly chosen member receives a higher attack score than a randomly chosen non-member, and is independent of any specific decision threshold. This property makes AUC a stable and reliable indicator of membership inference strength under ranking-based evaluation. As a complementary metric, we report Accuracy (ACC) (Choquette-Choo et al., 2021), which is computed under a fixed or selected threshold and reflects the proportion of samples correctly classified as members or non-members in a concrete decision setting.

Experimental Details. Unless specified otherwise, all evaluations are conducted under the following default setting: Llama-2-7B-chat is used as the generative module, E5-base is used as the retrieval module, Standard RAG is adopted as the RAG framework, and 4 retrieved documents are provided as context following prior work (Gao et al., 2025). Our experimental setup is identical to that of DCMI, including the same parameter configuration and the FlashRAG implementation. We successfully reproduce the DCMI results, which closely match those reported in the original paper. Therefore, under the shared setting we directly adopt the experimental numbers from the DCMI paper for comparison.

Table 7: Performance of ACC under different threshold selection strategies.

ACC	Llama			Mistral			Qwen		
	Fixed	GMM	Opt.	Fixed	GMM	Opt.	Fixed	GMM	Opt.
TREC-COVID	0.893	0.887	0.907	0.887	0.880	0.893	0.873	0.887	0.887
HealthCareMagic	0.893	0.953	0.953	0.927	0.933	0.947	0.847	0.900	0.900
MS MARCO	0.680	0.660	0.733	0.680	0.693	0.740	0.693	0.743	0.807

For QP-MIA, we set $\lambda_1 = 1.5$, $\lambda_2 = 2$, the fixed decision threshold $\tau = 13.5$, and the correctness threshold $\theta_{\text{cor}} = 0.8$. For each target document, we generate $m = 7$ fill-in-the-blank questions in a single exam and issue two composite queries of length $n = 5$. The two query sets share three overlapping questions and differ in the remaining four (two unique to each organization). The first query follows the original generation order, while the second applies a random permutation before concatenation.

C Additional Experimental Results

C.1 Additional Analysis

Effect of Query Composition Perturbation on Top-k Overlap.

To examine how query composition perturbation reshapes the retrieved evidence, we measure the size of the intersection between the top-k sets returned under the original and perturbed query compositions. As shown in Table 5, full overlap is uncommon for members and nearly absent for non-members on both TREC-COVID and HealthCareMagic. By contrast, zero or small intersections account for a substantially larger share of non-member cases—approximately 40%–50% under E5—than of member cases. Medium overlap remains frequent for members, indicating that a mild change in question content and presentation order is already sufficient to reshuffle the retrieved neighbors, while member targets more often retain a shared core of evidence across the two compositions. These patterns are consistent with the design motivation of QP-MIA: a genuine member document provides a stable shared evidence source that can remain effective under controlled dual perturbation, whereas a non-member target typically depends on fragmented matches to related documents and is more sensitive to changes in the retrieval context.

Table 5: Top-k intersection sizes under original and perturbed queries.

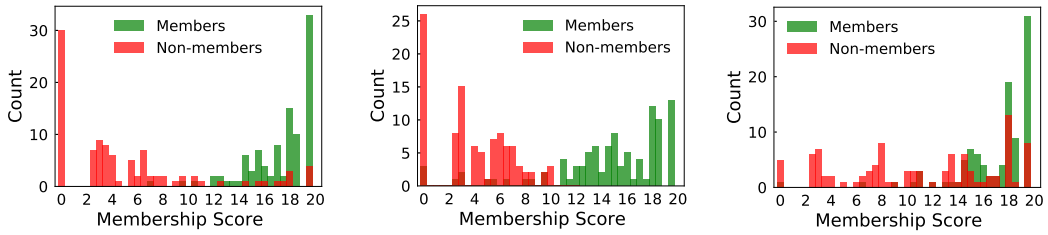
Count	TREC-COVID					HealthCareMagic				
	0	1	2	3	4	0	1	2	3	4
Member (E5)	0	28	28	31	13	15	31	33	18	3
Non-member	11	36	24	25	4	41	26	26	7	0
Member (Contriever)	1	15	48	28	8	15	39	24	20	2
Non-member	7	30	29	31	3	27	34	24	14	1

Effect of Original and Perturbed Queries on Target Retrieval.

To examine whether dual query composition probing provides a sufficient retrieval budget for genuine members, we measure how often the target document appears in the top-k under the original and perturbed query compositions. As shown in Table 6, the member target is retrieved at least once in almost all cases: under E5, the rate reaches 100% on TREC-COVID and approximately 91% on HealthCareMagic; under the weaker Contriever setting, it remains approximately 98% and 84%, respectively. Hits under both compositions are only slightly lower, while the single-composition rates already stay high. These results indicate that submitting two partially overlapping joint queries is sufficient to surface target-specific evidence for genuine members. Consequently, the subsequent contrastive membership signal—based on answer accuracy and cross-query stability—is not bottlenecked by the absence of the target document from the top-k.

Table 6: Target hit rate in top-k under original and perturbed queries.

Setting	TREC-COVID				HealthCareMagic			
	Orig.	Pert.	$\geq$1	Both	Orig.	Pert.	$\geq$1	Both
Member (E5)	1.00	0.99	1.00	0.99	0.85	0.84	0.91	0.78
Member (Con.)	0.98	0.97	0.98	0.97	0.70	0.79	0.84	0.65



(a) TREC-COVID (b) HealthCareMagic (c) MS-MARCO

Figure 4: Distribution of membership scores S for members and non-members under QP-MIA.

Effect of the Decision Threshold. Under the black-box setting, no labeled auxiliary data are available for tuning a decision boundary. When the attacker instead holds an unlabeled pool of documents under attack, a data-driven threshold can still be obtained via a Gaussian mixture model (GMM) without ground-truth membership labels. Specifically, the membership scores $\{S\}$ of this pool are first computed from the collected responses; a two-component GMM is then fitted to the empirical score distribution, under the assumption that member and non-member scores form two latent clusters. We take the intersection of the two fitted Gaussian densities as τ_{GMM} and predict documents with $S \geq \tau_{\mathrm{GMM}}$ as members and the remainder as non-members. This procedure uses only the score distribution of the attacked set and does not rely on any membership labels.

We next compare three ways of converting continuous scores into member/non-member decisions under an imbalanced evaluation pool with a non-member-to-member ratio of 2:1: a fixed threshold, a label-free two-component GMM threshold, and an oracle threshold that maximizes ACC using ground-truth labels. As shown in Table 7, the fixed threshold is already competitive and relatively stable across generators and datasets. The GMM threshold is typically close to the oracle, and on several settings matches the optimal ACC exactly. We further evaluate GMM as a practical alternative for threshold selection, because an attacker typically targets a set of candidates rather than a single document, so that a two-component mixture can be fitted on the batch of attack scores without a labeled validation set. We evaluate under a 2:1 non-member-to-member ratio to avoid the concern that performance is driven by a balanced mid-point decision rule, and to better reflect realistic deployments where non-members usually outnumber members.

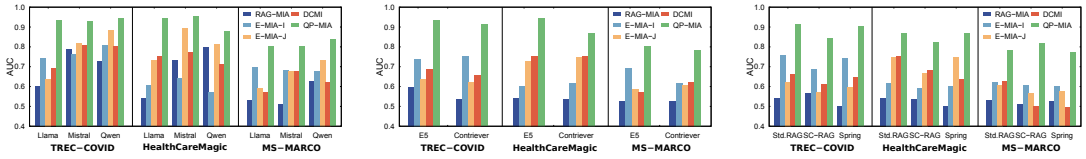
Effect of Cross-Query Correctness Consistency.

To further validate the source of the membership signal, we break down each target-specific question by how many times it is asked and how many times it is answered correctly under the dual partially overlapping design. As shown in Table 8, questions that are asked twice and answered correctly twice provide the strongest separation: the member-to-non-member count ratio reaches approximately $4.4\times$ on TREC-COVID and approximately $8.5\times$ on HealthCareMagic. In contrast, questions asked twice and answered correctly once are much weaker, with members only mildly outnumbering non-members on both datasets. These statistics support our scoring design, which places higher weight on repeated successful recovery of the same target-specific fact and treats partial recovery under only one composition as a weaker, less discriminative cue.

Table 8: Distribution of per-question correctness outcomes.

Count	TREC-COVID				HealthCareMagic			
	2✓2	2✓1	1✓1	0	2✓2	2✓1	1✓1	0
Member	229	53	336	82	170	75	272	183
Non-member	52	45	91	512	20	47	76	557

Distribution of Membership Scores. Figure 4 shows the distribution of the membership score S for members and non-members across the three datasets. Taking HealthCareMagic as a representative example, member scores concentrate almost exclusively in the high range, whereas non-member scores are heavily skewed toward the low end. Consequently, the high-score bins become nearly pure, with member ratios reaching 100%. Similar separation patterns are observed on TREC-COVID and MS-MARCO, although the latter exhibits a modest degree of overlap in the mid-range. This clear polarization indicates that the dual partially overlapping query design



(a) Generators (b) Retrievers (c) RAG Frameworks

Figure 5: AUC under different generators, retrievers, and RAG frameworks.

of QP-MIA induces a robust contrast in answer accuracy and cross-query stability that effectively separates members from non-members. The sharpest contrast—and correspondingly the strongest attack performance—is achieved on HealthCareMagic, further confirming the effectiveness of the proposed scoring mechanism.

Visual Comparison under Varying Configurations. To provide a more intuitive comparison of performance across the three datasets under different generators, retrievers, and RAG frameworks, we present the corresponding bar charts in Figure 5. QP-MIA (shown in green) consistently achieves the highest AUC in all settings, clearly demonstrating its superiority over the baseline methods.

C.2 Evaluation against Potential Defenses

We further examine the robustness of QP-MIA under representative defense strategies against RAG membership inference. Specifically, we consider defenses that constrain the generation process, rewrite the submitted questions, or transform the exposed context after retrieval. For each defense, we evaluate QP-MIA on HealthCareMagic and compare it with existing attacks to analyze how different protective mechanisms affect membership leakage.

Instruction-Based Defense (Instr.). Our first defense augments the RAG system prompt with explicit safety instructions. The added instructions require the model to (i) refuse any request that appears to probe the underlying corpus, (ii) avoid repeating or quoting retrieved passages in raw form, and (iii) answer only on the basis of general knowledge when the query is suspected to be membership-oriented. This strategy only reshapes the model's response policy at the prompt level and does not alter retrieval or document content. Empirically, the resulting protection is limited. As shown in Table 4, QP-MIA remains essentially unchanged, whereas existing methods exhibit only limited degradation. These results indicate that simple instruction-level constraints are insufficient to suppress the membership signal that QP-MIA derives from answer accuracy and cross-query stability.

Paraphrasing-Based Defense (Paraph.). The second defense rewrites each user query into an alternative form while preserving its semantic intent. In our implementation, an auxiliary LLM rephrases the original fill-in-the-blank questions into synonymous but lexically different expressions before they are submitted to the RAG system, thereby reducing lexical overlap with stored documents. As shown in Table 4, this strategy is substantially stronger than prompt constraints: DCMI falls sharply from 0.75 to 0.56. In contrast, QP-MIA declines only moderately to 0.88 and remains clearly ahead of all baselines. This robustness stems from dual content-and-order perturbation with partially overlapping joint queries: because membership is inferred from aggregate answer accuracy and cross-query stability rather than from a single surface form, lexical paraphrasing only partially weakens the signal and does not fully eliminate it.

Post-retrieval Entity-relation Extraction (Entity-Rel.). The third defense operates after retrieval and before generation by converting the retrieved documents into entity–relation triples. In practice, a lightweight information-extraction module is applied to each retrieved passage, retaining only structured triples of the form (entity, relation, entity) while discarding the original free-text surface. This transformation removes much of the lexical evidence that attacks can exploit, while still allowing the generator to synthesize answers from the structured knowledge. In our evaluation, it provides the strongest mitigation among the three defenses. DCMI is reduced to near-random performance, and the E-MIA variants degrade substantially. QP-MIA decreases to 0.79, yet still outperforms all baselines by a clear margin. Even under this stronger intervention, QP-MIA contin-

ues to retain a usable membership signal from answer accuracy and cross-query stability under dual query composition perturbation.

Overall, these results reveal a clear gap between lightweight prompt interventions and stronger content-transformation defenses. Constraining response behavior at the instruction level offers only limited protection, while rewriting queries or restructuring retrieved evidence yields progressively stronger mitigation. Even under the strongest defense, however, QP-MIA retains a substantial advantage over existing attacks, indicating that membership signals arising from dual-perturbation answer accuracy and cross-query stability are difficult to eliminate completely in practical RAG deployments.