E-MIA: EXAM-STYLE BLACK-BOX MEMBERSHIP INFERENCE ATTACKS AGAINST RAG SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) equips large language models (LLMs) with external evidence by retrieving documents at inference time, but it also turns the retrieval corpus into a sensitive asset. Under a black-box setting, an adversary *given a candidate document* can infer whether it has been ingested into the RAG knowledge base (*i.e.*, document-level membership inference) solely from query–response interactions, thereby leaking corpus coverage and the existence of sensitive topics. Existing RAG Membership Inference Attack (MIA) methods either rely on "soft" signals such as semantic similarity, which often yield overlapping member/non-member score distributions and unstable thresholds, or employ explicit confirmation probes whose intent is conspicuous and thus prone to refusal and detection. We propose E-MIA, which converts verifiable hard evidence in the target document (*e.g.*, fine-grained details, proper nouns/technical terms, definitional statements, metadata cues, and causal/constraint relations) into an exam with four objectively gradable question types (FB/SC/MC/T/F), and uses the aggregated exam score across multiple evidence-targeted questions as the membership signal. Experiments across multiple datasets and diverse RAG configurations demonstrate that E-MIA improves member/non-member separability in stringent settings while preserving natural, stealthy queries, and we further analyze the impact of question composition and exam length on attack effectiveness.

1 INTRODUCTION

Retrieval-augmented generation (RAG) conditions large language model (LLM) outputs on documents retrieved from an external corpus, allowing knowledge-intensive systems to use updatable evidence without retraining Lewis et al. (2020); Borgeaud et al. (2022); Gao et al. (2024). The same corpus, however, becomes a sensitive asset. We study *document-level membership inference* against RAG systems: given a candidate document, a black-box adversary determines whether it belongs to the retrieval corpus using only query–response interactions. Unlike membership inference over model training data or parametric memorization Shokri et al. (2017); Meeus et al. (2024), the target is corpus inclusion; even this binary disclosure may reveal the coverage of internal policies, business records, or copyright-protected sources.

Existing black-box RAG membership inference attacks derive evidence from response–target similarity, masked completion, document-derived confirmation questions, differential calibration, or generation-budget side channels Li et al. (2025c); Anderson et al. (2025); Liu et al. (2025); Naseh et al. (2025); Gao et al. (2025); Li et al. (2025a); Feng et al. (2025); WU et al. (2024). These attacks establish that corpus membership can leak through model outputs, but content-based signals remain subject to two opposing confounders: a query for a member document may fail to retrieve it, whereas a query for a non-member may still be answered from parametric knowledge, related corpus documents, or guessing. Similarity scores may therefore reward plausible paraphrases, while explicit confirmation-style questions can expose the probing intent. The challenge is not simply to issue more queries, but to construct repeated observations that are document-specific, objectively gradable, and naturally expressed under a black-box interface.

Our key insight is to shift the unit of inference from overall response similarity to *verifiable hard evidence* in the candidate document. Each evidence point should admit deterministic grading, discriminate the candidate from generic or overlapping knowledge, and be expressible as a conven-

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.

tional question. Precise details, proper names and technical terms, definitions, metadata cues, and constraint relations often satisfy these requirements. When such evidence is retrievable but not readily available from parametric knowledge or other documents, consistent performance across distinct evidence points provides a membership signal; when the evidence is common or duplicated, that signal necessarily weakens.

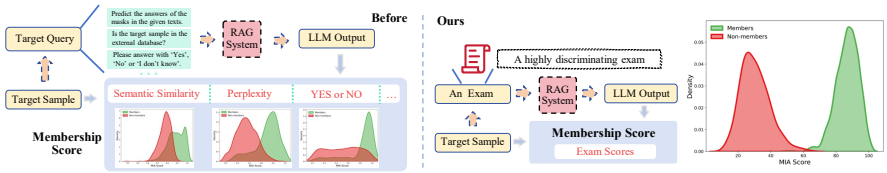


Figure 1: Conceptual comparison of membership signals in black-box RAG systems. Representative prior attacks use output-level similarity, perplexity, or confirmation signals. E-MIA instead queries multiple document-specific evidence points and aggregates objectively graded answers into an exam score.

Building on this insight, we propose **E-MIA** (Exam-style Membership Inference Attack), which converts typed evidence units into four objectively graded question formats—single-choice (SC), multiple-choice (MC), true/false (T/F), and fill-in-the-blank (FB)—and aggregates the target system's performance into a document-level membership score. Its design links each challenge to a concrete mechanism: canonical answers replace ambiguous global similarity with evidence-level checks; conventional exam questions avoid explicit corpus-membership prompts; and aggregation across multiple evidence points and formats limits the influence of any single retrieval miss, accidental hit, or plausible generic answer. At inference time, E-MIA uses only final textual outputs and does not observe the knowledge base, retriever scores, retrieved contexts, or internal logs. Fig. 1 summarizes this shift from output-level signals to aggregated evidence verification. Our source code is released in `https://anonymous.4open.science/r/EMIA-SADAQSDFDQDWQED80`

We summarize our **contributions** as follows:

- We formulate black-box RAG document membership inference around typed, verifiable evidence units with canonical answers and deterministic grading rules.
- We develop E-MIA, an exam-style attack that maps these units to four conventional question formats and aggregates their graded outputs into a document-level membership score using only textual API interactions.
- We show that, holding exam length, type mix, grading rule, and aggregation fixed, the choice of evidence source alone shifts membership separability by up to 0.538, and we evaluate E-MIA against three RAG membership inference baselines across three corpora and four target large language modes (LLMs), including under query/response rewriting and malicious-input detectors.

2 Related Work

Retrieval-Augmented Generation (RAG) grounds generation in evidence retrieved from an external corpus Ji et al. (2023); Gao et al. (2024), which in turn makes that corpus a privacy asset: an adversary may infer, from black-box outputs alone, whether a given document is retrievable Liu et al. (2025). The constraint that makes this problem hard is that the adversary observes only the final text—not the index, the retriever scores, the retrieved contexts, or any internal logs—while the queries themselves must still look like ordinary use. Existing RAG-MIA methods differ mainly in where they draw the membership signal. Each class pays for its choice. (i) *Explicit confirmation* asks the system directly whether target content appears in the retrieved context; the signal is strong, but the intent is stated in the query itself and is therefore easily refused or flagged by the guardrails and prompt-injection detectors that RAG APIs now deploy Anderson et al. (2025). (ii) *Similarity- or quality-based scoring* derives the score from semantic similarity, perplexity, or answer quality; such signals remain black-box friendly, but they measure how familiar the generator already is with the topic rather than whether the corpus supplied it, so a topically related non-member is answered just as fluently as a member and the two score distributions overlap Li et al. (2025c); Naseh et al. (2025). (iii) *Content probing* amplifies the member/non-member gap through masked completion, perturbation calibration, or side channels (*e.g.*, generation budgets), at the cost of stronger assumptions and more complex pipelines Gao et al. (2025); Liu et al. (2025); Li et al. (2025a); Wang et al.

(2025a). The underlying trade-off is thus between robustness and feasibility: soft signals give overlapping score distributions and unstable thresholds, whereas enhancement strategies introduce extra assumptions and are more likely to trigger defenses Varshney et al. (2024). Our E-MIA departs from both by grounding membership in evidence intrinsic to the document and casting the task as an exam-style evaluation whose aggregated score is the membership signal. Appendix A.2 provides a fuller discussion of RAG and RAG-MIA.

3 Threat Model

Attacker's Target. We study document-level membership inference in RAG systems. Given a target document d^*, the adversary aims to determine whether d^* is contained in the RAG knowledge base $\mathcal{D}$, *i.e.*, whether $d^* \in \mathcal{D}$ holds.

Attacker's Constraints. We consider a realistic black-box setting: the adversary has no access to $\mathcal{D}$, retriever/LLM parameters, retrieved contexts $\mathcal{P}_k$, or retrieval scores, and may be filtered by a malicious-input detector. The adversary can only issue queries and observe final textual outputs. We assume the adversary possesses the content of d^* (as standard in membership inference), but cannot access the system's knowledge base or retrieval traces. To improve feasibility and stealth, we craft prompts as natural Questions/Answer (QA) or reading-comprehension questions rather than explicit confirmation-style probes.

Attacker's Task. We instantiate the attack as an exam-style procedure. The adversary generates a single exam $X_{d^*} = \{(q_i, a_i)\}_{i=1}^n$, where q_i is an exam question derived from evidence in d^* and a_i is its ground-truth answer. Submitting the entire exam as a single query to the RAG system yields responses $\{r_i\}_{i=1}^n$, and an *Exam Evaluator* grades each response and aggregates the results into an exam score:

$$S(d^*) = \mathrm{Agg}(\{\mathrm{Score}(r_i, a_i)\}_{i=1}^n), \tag{1}$$

which Section 4 instantiates as a weighted aggregation of type-wise accuracies. Finally, the adversary outputs a membership decision $\hat{m}(d^*) = \mathbb{I}[S(d^*) \geq \tau]$. In our design, $\mathrm{Score}(\cdot)$ is deterministic and objectively checkable, and aggregation over multiple questions mitigates variance from retrieval mismatch.

4 E-MIA Methodology

This section presents E-MIA, which reformulates document-level membership inference as an exam-style verification of evidence extracted from a target document. Specifically, as illustrated in Fig. 2, given an evidence set $\mathbf{E}(d^*)$, we construct natural-language queries and generate a structured set of questions; the system's performance on this question set is then evaluated via objective scoring to produce a stable membership inference score. We provide detailed examples and prompt templates for our method in Appendix A.4–A.6.

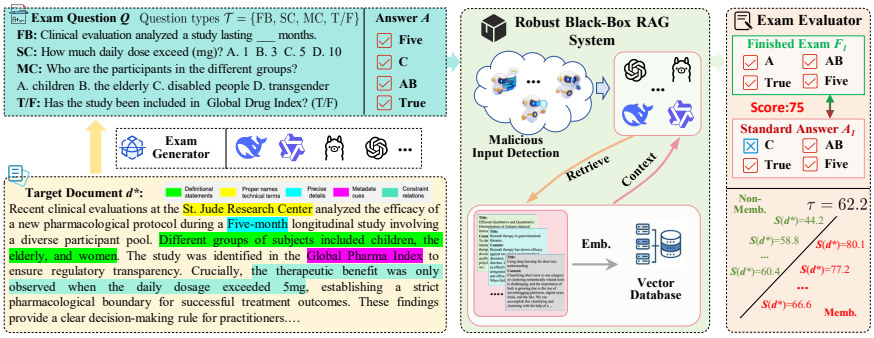


Figure 2: Overview of E-MIA, which infers document membership in a black-box RAG system by generating an exam from a target document, querying the RAG to complete it, and grading the answers to obtain a membership score

4.1 Evidence-Centric Document Decomposition

In a black-box RAG system, the adversary observes only final responses, and signals built on global semantic similarity or coarse answer quality are unstable: even when d^* is not retrieved, the model

may still produce plausible answers through generalization and paraphrasing, causing member and non-member score distributions to overlap Meeus et al. (2024); Yang et al. (2026); Gao et al. (2025).

We therefore shift the inference focus from overall similarity to whether the output consistently matches document-specific, verifiable hard evidence: an ingested document becomes retrievable with non-trivial probability, so the system is more likely to reproduce its key facts and constraints across multiple queries, whereas deviations on those points occur more frequently otherwise.

To balance measurability with stealthiness, we require each evidence point to satisfy three criteria: it must admit objective grading, be discriminative of the target document, and be naturally expressible as conventional QA Mansurova et al. (2024) or reading-comprehension style questions. Under these criteria, the most discriminative evidence tends to fall into five categories:

- **Precise details**: numbers, dates, thresholds, ordering, and conditional clauses, *et al.*
- **Proper names and technical terms**: entities, acronyms, and domain-specific terminology, *et al.*
- **Definitional statements**: crisp concept boundaries and exact definitions, *et al.*
- **Metadata cues**: titles, section/clause identifiers, versions, and other structural markers, *et al.*
- **Constraint relations**: premise–conclusion, condition–action, and dependency links, *et al.*

These categories are not chosen for maximum coverage; rather, they are hard to guess reliably from "generalization-only" answers while remaining easy to grade deterministically.

We represent each target document d^* by an evidence set $\mathbf{E}(d^*) = \{e_j\}_{j=1}^{M}$, where each evidence unit is represented as $e_j = (c_j, y_j, t_j, g_j)$, where, c_j denotes the evidence content or semantic anchor, y_j is its canonical answer, $t_j \in \mathcal{T}$ specifies the question type, and g_j denotes the corresponding deterministic grading rule. The question-type set is

$$\mathcal{T} = \{\mathrm{FB}, \mathrm{SC}, \mathrm{MC}, \mathrm{T/F}\}. \tag{2}$$

From these evidence units, we derive objectively gradable exam items while preserving the evidence content and its canonical answer. This representation explicitly separates the document-specific evidence from its natural-language realization, allowing the same evidence unit to be instantiated into different question forms while retaining a common grading target.

4.2 Exam Query Construction

Given $\mathbf{E}(d^*)$, we construct a pool of candidate exam items $\mathcal{Q}$ by treating each evidence unit e_j as the minimum generation granularity. For each evidence unit, a question is instantiated through an evidence-to-question mapping

$$q_j = \phi(e_j), \qquad a_j = y_j, \tag{3}$$

where $\phi(\cdot)$ generates a natural-language question from the corresponding evidence unit. The resulting item is therefore represented as

$$x_j = (q_j, a_j, t_j, g_j). \tag{4}$$

We retain essential metadata—question type, prompt, answer, and scoring rule—for consistent grading from textual outputs alone. To enhance stealthiness, prompts mimic conventional reading-comprehension or fact-verification questions, avoiding explicit confirmation. Following minimal exposure, we reveal only short anchors (*e.g.*, entities or keywords) rather than verbatim passages, and support limited output variations via deterministic normalization and pattern extraction (*e.g.*, option-letter extraction, numeric/date canonicalization).

Question instantiation follows an evidence-to-question mapping, calibrated offline (Section 5.3, Fig. 5). Proper names, technical terms, and local causal/constraint relations become true/false items testing a single entity or logical link directionality. Precise details, definitional statements, and metadata cues become fill-in-the-blank items, preserving numbers, concept boundaries, and structural markers (titles, clause identifiers, versions). Precise details also supply attributes for multiple-choice items, aggregating several facts into one multi-attribute item. To mitigate accidental hits and generalization noise, the same evidence unit is often measured redundantly via multiple paraphrases and/or formats, yielding a more stable signal. We then filter out questions with ambiguous answers,

non-objective grading, or insufficient stealthiness, and normalize answer representations (*e.g.*, units, date formats, alias sets) to ensure consistent scoring and facilitate exam assembly and score aggregation. As illustrated in Fig. 3, we instantiate a target document into four objectively-gradable item types (FB, SC, MC, and T/F) while keeping prompts in a natural reading-comprehension style.

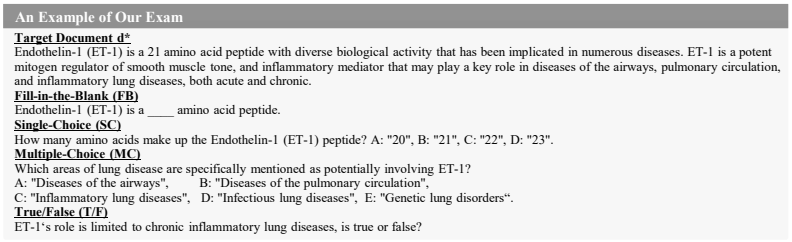


Figure 3: Example of an exam automatically generated from a target document d^* (Endothelin-1), including four question types: Fill-in-the-Blank, Single-Choice, Multiple-Choice, and True/False.

4.3 Exam Generation

Given the candidate exam pool $\mathcal{Q}$, exam generation follows a three-stage pipeline: question-type specification, assembly, and validation. We use membership separability, $\Delta = Acc_{\text{mem}} - Acc_{\text{non}}$, only for offline calibration on a held-out development set to guide design choices; the attack itself uses fixed specifications.

Based on this calibration (see Section 5.3 for mechanism analysis), we adopt the following default, reproducible rules for each question type:

- **Fill-in-the-Blank (FB):** We use one or two key blanks by default, which preserves minimal context while enforcing strong constraints.
- **Single-Choice (SC):** We use a small option set with a limited number of semantically similar distractors; specifically, $C \in \{4, 5\}$ total options and $D \in \{1, 2\}$ similar distractors.
- **Multiple-Choice (MC):** We use $n \in \{4, 5\}$ options with the number of correct answers fixed to $m = 2$.
- **True/False (T/F):** We generate items using fixed templates and consistent phrasing constraints to keep grading simple and stable.

After fixing type-specific specifications, we bucket questions by evidence category and question type, and sample N questions (typically a multiple of 4 to control the type mix). The sampling strategy prioritizes coverage of diverse evidence points, while allowing only a small subset of critical evidence units to be measured redundantly via paraphrases or cross-type variants to reduce variance from accidental hits and generalization noise. We then perform consistency checks and cleaning, removing questions with multiple valid answers, those requiring overly long context for objective grading, and prompts whose wording may reveal explicit probing intent. For choice-based items, distractors are constructed from nearby-but-incorrect candidates (*e.g.*, close numerical values, similar entities, condition substitutions, or reversed causal directions) to avoid overly vague decision boundaries. For FB questions, we enforce deterministic answer-format and normalization rules (*e.g.*, units, date formats, alias sets) to ensure consistent grading based solely on the model output. Through this assembly-and-validation procedure, we obtain a complete exam that is natural in appearance, objectively scorable, and reliably discriminative for subsequent inference scoring and threshold-based membership decisions.

4.4 Inference Attack Scores

To quantify membership, we normalize the exam outcome to a 0–100 *membership score*. Let $\mathcal{T} = \{\text{FB}, \text{SC}, \text{MC}, \text{T/F}\}$ denote the four question types. For each type t, let $X_t = \{j : t_j = t\}$ and $N_t = |X_t|$ be the set and number of exam items of that type. For each exam item j, let r_j denote the RAG response and define its binary grading outcome as $Z_j = \text{Score}(r_j, a_j) \in \{0, 1\}$, where $Z_j = 1$ indicates that the response satisfies the objective grading criterion associated with the target evidence. The type-wise empirical correctness rate is $\widehat{p}_t = \frac{1}{N_t} \sum_{j \in X_t} Z_j$.

We aggregate the type-wise correctness rates using a fixed weighted sum:

$$S(d^*) = 100 \sum_{t \in \mathcal{T}} w_t \widehat{p}_t, \tag{5}$$

where $w_t \geq 0$ and $\sum_{t \in \mathcal{T}} w_t = 1$.

Let M denote the membership status, with $M = 1$ if $d^* \in \mathcal{D}$ and $M = 0$ otherwise. For a question type t, we define its conditional correctness probability as $p_t^{(m)} = \Pr(Z = 1 \mid M = m, t_j = t)$, for $m \in \{0, 1\}$. For document-specific evidence, E-MIA relies on the empirical premise that

$$p_t^{(1)} > p_t^{(0)}, \tag{6}$$

i.e., the target document is expected to induce a higher probability of correctly reproducing its specific evidence when it is present in the knowledge base. Eq. (6) also gives a direct definition of the membership separability used for offline analysis. Let $\bar{S} = S/100$ and define $\mu_m = \mathbb{E}[\bar{S} \mid M = m]$. By construction,

$$\Delta = \mu_1 - \mu_0 = \sum_{t \in \mathcal{T}} w_t \left(p_t^{(1)} - p_t^{(0)} \right), \tag{7}$$

where Δ measures the expected score separation between members and non-members. This decomposition shows that the overall membership signal is determined by the weighted contribution of each question type.

We choose w_t via offline calibration on a held-out development set by emphasizing question types whose member/non-member score distributions are more separable (measured by Kullback-Leibler (KL) divergence), and keep w_t fixed during attack. We use the calibrated default weights reported in Section 5.3. Unless otherwise specified, we use the calibrated defaults $w_{\mathrm{FB}} = 0.312$, $w_{\mathrm{SC}} = 0.214$, $w_{\mathrm{MC}} = 0.300$, and $w_{\mathrm{T/F}} = 0.174$.

Under conditional independence of question-level grading outcomes within each membership state, the aggregation also yields

$$\mathrm{Var}(\bar{S} \mid M = m) = \sum_{t \in \mathcal{T}} \frac{w_t^2 p_t^{(m)} \left(1 - p_t^{(m)}\right)}{N_t}. \tag{8}$$

This relation motivates the use of multiple evidence-targeted questions: increasing N_t reduces the sampling variance of the aggregated score, while the type-wise separation in Eq. (7) determines the magnitude of the membership signal.

Finally, we output a binary membership decision with a threshold: $\hat{m}(d^*) = \mathbb{I}[S(d^*) \geq \tau]$, where we select τ on the same calibration set to maximize balanced accuracy (*cf.* Fig. 7), and use the default $\tau = 62.2$ in all experiments.

5 Experiments

In this section, we empirically evaluate whether E-MIA can infer document membership in black-box RAG systems across three corpora, four target LLMs, and three retrievers. We first describe the experimental setup (Section 5.1), followed by a comparison against three RAG-MIA baselines under stringent false-positive constraints (Section 5.2). We then analyze why the exam-style design works, covering question structure, evidence-to-question mapping, and calibration sensitivity (Section 5.3), and examine its robustness against rewriting defenses, malicious-input guardrails, and varying retrieval budgets (Section 5.4).

5.1 Experimental Setup

Datasets. We evaluate on three BEIR corpora Thakur et al. (2021) covering biomedical and scientific text: NFCorpus, TREC-COVID, and SCIDOCS (3.5K, 23K, and 116K documents). We designate 60% of each corpus as members, which are ingested into the RAG knowledge base, and hold out the remaining 40% as non-members, from which we draw a balanced evaluation set of

1,000 members and 1,000 non-members per dataset. Because non-members are never ingested, any correct answer on their evidence must come from memorization or spurious correlation rather than from direct retrieval. Appendix A.3.1 reports the corpus statistics, the preprocessing pipeline, and the sampling procedure.

Exam generator and evaluator. We generate exams with the DeepSeek-R1 Guo et al. (2025) API using question-type-specific prompt templates (Appendix A.5.1). Grading is fully objective: after normalization, FB and T/F items are scored by exact match, while MC and SC items are scored by extracting option identifiers and comparing them as sets (Appendix A.6).

Target RAG systems. We instantiate a standard RAG pipeline with four generators of different scales—Gemma-2-2B Team et al. (2024), Qwen-2.5-7B Yang et al. (2024), Llama-3.1-8B Dubey et al. (2024), and DeepSeek-R1-14B Guo et al. (2025)—and three retrievers spanning different embedding spaces, BGE Zhang et al. (2026), MiniLM-L6-v2 Wang et al. (2020), and text-embedding-3-small Zhang et al. (2025). We further equip the pipeline with a security layer that combines query and response rewriting Ma et al. (2023); Dharwada et al. (2025); Wang et al. (2025b) with three guardrail models: PIGuard Li et al. (2025b), Tanaos-Guardrail-v1, and FT-Llama-Prompt-Guard-2 Security (2024) (Appendix A.7). Unless stated otherwise, we use k=3, BGE, and N=28 exam questions.

Baselines and metrics. We compare against three RAG-MIA baselines—S^{2}MIA Li et al. (2025c), MBA Liu et al. (2025), and IA Naseh et al. (2025)—and report Accuracy, AUC-ROC, AUC-PR, and TPR@FPR (0.05/0.01/0.005); Appendix A.3 describes each baseline and metric.

5.2 Attack Effectiveness

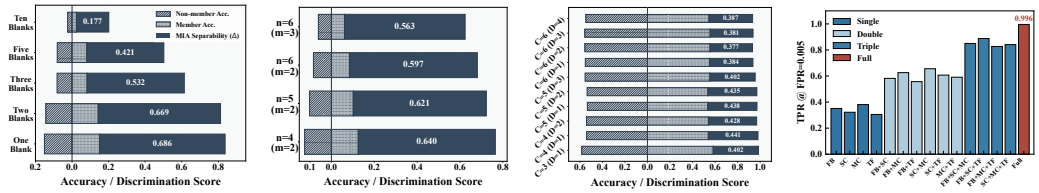
Table 1 compares E-MIA with the three baselines across three corpora and four LLM backbones. E-MIA is consistently near-perfect (Accuracy above 0.97, AUC-ROC up to 1.000), and it stays stable exactly where the baselines collapse: at FPR$=0.005$ their TPR frequently drops below 0.40, whereas E-MIA never falls below 0.817 and reaches 0.998. Averaged over backbones, this corresponds to a 117.8–142.6% improvement in TPR@FPR$=0.005$ over the strongest baseline. The gap indicates that E-MIA recovers document-specific evidence that similarity-based and completion-based black-box attacks fail to expose.

Table 1: Comparison of membership inference attacks across three RAG corpora and four LLM backbones (Deepseek-R1-14B, Llama3.1-8B, Qwen2.5-7B, and Gemma2-2B), reporting Acc, AUC-ROC, AUC-PR, and TPR@FPR (0.05/0.01/0.005).

	Deepseek-r1-14B				Llama3.1-8B				Qwen2.5-7B				Gemma2-2B			
Metric	**S^{2}MIA**	**MBA**	**IA**	**E-MIA**	**S^{2}MIA**	**MBA**	**IA**	**E-MIA**	**S^{2}MIA**	**MBA**	**IA**	**E-MIA**	**S^{2}MIA**	**MBA**	**IA**	**E-MIA**
NFCorpus																
Accuracy	0.714	0.726	0.925	**0.977**	0.676	0.734	0.932	**0.995**	0.741	0.757	0.932	**0.990**	0.660	0.627	0.929	**0.990**
AUC-ROC	0.796	0.805	0.914	**0.995**	0.802	0.842	0.929	**0.996**	0.834	0.835	0.928	**0.982**	0.773	0.723	0.926	**0.997**
AUC-PR	0.833	0.841	0.915	**0.994**	0.838	0.864	0.913	**0.989**	0.868	0.878	0.932	**0.986**	0.814	0.711	0.921	**0.994**
TPR@0.05	0.487	0.468	0.816	**0.993**	0.485	0.558	0.967	**1.000**	0.555	0.558	0.917	**0.985**	0.454	0.357	0.947	**1.000**
TPR@0.01	0.357	0.354	0.523	**0.911**	0.365	0.432	0.827	**0.998**	0.447	0.455	0.752	**0.984**	0.330	0.254	0.612	**0.988**
TPR@0.005	0.314	0.309	0.254	**0.817**	0.332	0.338	0.505	**0.976**	0.387	0.322	0.474	**0.984**	0.291	0.229	0.401	**0.842**
TREC-COVID																
Accuracy	0.749	0.785	0.876	**0.989**	0.732	0.745	0.915	**0.991**	0.767	0.761	0.942	**0.996**	0.703	0.687	0.932	**0.992**
AUC-ROC	0.829	0.856	0.915	**0.998**	0.833	0.842	0.952	**0.999**	0.845	0.838	0.938	**1.000**	0.792	0.756	0.922	**0.999**
AUC-PR	0.865	0.869	0.895	**0.997**	0.869	0.852	0.935	**0.999**	0.885	0.883	0.956	**1.000**	0.833	0.784	0.955	**0.999**
TPR@0.05	0.546	0.568	0.877	**0.999**	0.557	0.574	0.897	**0.999**	0.611	0.456	0.899	**1.000**	0.486	0.417	0.827	**0.998**
TPR@0.01	0.528	0.514	0.504	**0.997**	0.461	0.472	0.416	**0.992**	0.518	0.434	0.607	**0.997**	0.352	0.302	0.533	**0.991**
TPR@0.005	0.377	0.362	0.208	**0.837**	0.393	0.373	0.393	**0.982**	0.462	0.353	0.406	**0.996**	0.303	0.218	0.356	**0.988**
SCIDOCS																
Accuracy	0.759	0.762	0.935	**0.985**	0.746	0.758	0.950	**0.997**	0.782	0.779	0.896	**0.999**	0.722	0.640	0.931	**0.996**
AUC-ROC	0.834	0.808	0.950	**0.998**	0.832	0.834	0.944	**0.999**	0.859	0.862	0.923	**0.999**	0.803	0.678	0.912	**0.999**
AUC-PR	0.864	0.824	0.958	**0.999**	0.864	0.876	0.956	**0.999**	0.890	0.896	0.937	**1.000**	0.839	0.709	0.915	**0.999**
TPR@0.05	0.510	0.506	0.923	**0.997**	0.523	0.592	0.927	**0.999**	0.596	0.595	0.788	**0.999**	0.474	0.352	0.920	**0.999**
TPR@0.01	0.366	0.426	0.684	**0.976**	0.395	0.403	0.405	**0.998**	0.487	0.471	0.393	**0.998**	0.352	0.277	0.695	**0.997**
TPR@0.005	0.338	0.361	0.397	**0.954**	0.345	0.384	0.392	**0.997**	0.434	0.432	0.393	**0.998**	0.334	0.248	0.405	**0.996**

5.3 Mechanism and Ablations

We now examine why E-MIA works, considering the question structure, the evidence-to-question mapping, the resulting score distributions, and the sensitivity to the offline calibration parameters.



(a) Different blank counts in FB (b) Option design factors in MC (c) Option design factors in SC (d) Question type combination

Figure 4: Effectiveness of the question types.

Structural Effectiveness. Fig. 4(a)–(c) reports the membership separability Δ of individual question parameters. SC is most discriminative with $C \in \{4, 5\}$ options and $D \in \{1, 2\}$ similar distractors, FB with one or two blanks, and MC with $n \in \{4, 5\}$ options carrying $m = 2$ correct answers; adding further blanks makes items so hard that even a system holding the target document answers them incorrectly. Fig. 4(d) shows that format diversity amplifies the signal further: mixing all four types reaches a peak TPR@FPR$= 0.005$ of 0.996 and outperforms every single-format configuration. The heterogeneous exam therefore exploits several facets of memorization rather than one.

Impact of evidence-question matching. Fig. 5 motivates a bidirectional priority between evidence and question types. Proper names and technical terms (PNT) and constraint relations (CR) are routed to T/F items, where the model's binary certainty about a specific entity or logical link is sharpest; precise details (PD), definitional statements (DS), and metadata cues (MDC) fill FB items, which demand faithful reproduction of numbers, concept boundaries, and structural markers. The reverse preference is consistent with this mapping: PNT and CR are the primary sources for T/F items, reaching a peak Δ of 0.790, while PD and DS dominate FB items and PD also feeds MC items, where several facts are aggregated into a single multi-attribute item. This dual mapping queries each factual anchor in the format that maximizes Δ.

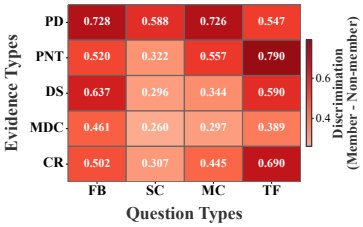
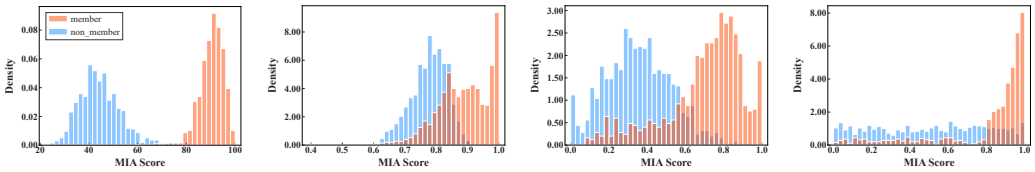


Figure 5: Heatmap of membership separability (Δ)

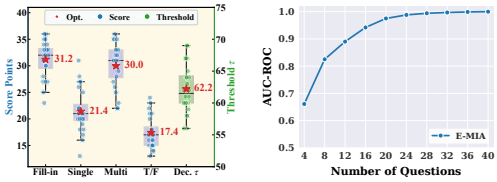
Distributional Advantage. Fig. 6 plots the score distributions on TREC-COVID. E-MIA (Fig. 6(a)) produces a markedly wider bi-modal gap between members (red) and non-members (blue), whereas S^{2}MIA (Fig. 6(b)), MBA (Fig. 6(c)), and IA (Fig. 6(d)) overlap substantially and hence admit more false positives. With little density left in the overlap region, high-confidence decisions rest on verifiable document evidence rather than on ambiguous surface similarity.



(a) E-MIA (ours) (b) S^{2}MIA (c) MBA (d) IA

Figure 6: Membership and non-membership score distributions.

Hyperparameter Sensitivity and Exam Length. E-MIA never observes the target system: the type weights w_i, the decision threshold τ, and the exam length N are fixed offline before any query is issued, so we test how sensitive the attack is to them. We calibrate w_i by prioritizing the most separable question types and τ by maximizing balanced accuracy, over 25 trials spanning the three corpora and four target LLMs. Instead of the per-configuration optimum we report the



(a) Optimal Parameters (b) Query Scalability

Figure 7: Impact of Question Density and Design

medoid configuration (red stars in Fig. 7(a)), which avoids overfitting a single deployment and is the appropriate choice under a black-box threat model; this yields the fixed allocation FB (31.2), SC (21.4), MC (30.0), and T/F (17.4) with $\tau = 62.2$. Exam length is the parameter that also drives query cost: Fig. 7(b) shows AUC-ROC rising steeply and saturating at roughly 20–24 questions, so we default to $N = 28$ (at least seven items per type), a stable operating point at small marginal cost.

5.4 Robustness and Stealth

We evaluate E-MIA under RAG security defenses and across retrieval configurations, the two conditions that most directly determine its reliability in deployment.

Table 2: Robustness and stealthiness of the compared attacks. (a) Accuracy under query, response, and combined rewriting; retrieval recall is analysis-only and not attacker-observable, and parenthetical values denote the degradation relative to the undefended setting. (b) Fraction of queries that bypass each malicious-input detector; higher is stealthier.

(a) Robustness against rewriting				(b) Stealthiness against detectors		
Rewriting Strategy	**Method**	**Retrieval Recall**	**Accuracy**	**Detection Model**	**Method**	**Pass Rate**
Query Rewriting	S^2MIA	0.942 (0.041↓)	0.722 (0.019↓)	**PIGuard**	S^2MIA	0.439
	MBA	0.971 (0.011↓)	0.597 (0.010↓)		MBA	0.533
	IA	0.871 (0.115↓)	0.801 (0.181↓)		IA	0.976
	E-MIA(Ours)	**1.000 (0.000↓)**	**0.988 (0.002↓)**		E-MIA(Ours)	**1.000**
Response Rewriting	S^2MIA	0.982 (0.001↓)	0.484 (0.257↓)	**Tanaos-Guardrail-v1**	S^2MIA	0.984
	MBA	0.979 (0.003↓)	0.601 (0.006↓)		MBA	0.653
	IA	0.986 (0.000↓)	0.963 (0.019↓)		IA	0.991
	E-MIA(Ours)	**1.000 (0.000↓)**	**0.987 (0.003↓)**		E-MIA(Ours)	**0.997**
Both Rewriting	S^2MIA	0.950 (0.033↓)	0.462 (0.279↓)	**FT-Llama-Prompt-Guard-2**	S^2MIA	0.210
	MBA	0.972 (0.010↓)	0.598 (0.009↓)		MBA	0.456
	IA	0.825 (0.161↓)	0.691 (0.291↓)		IA	0.992
	E-MIA(Ours)	**1.000 (0.000↓)**	**0.988 (0.002↓)**		E-MIA(Ours)	**1.000**

Robustness to rewriting defenses. Across query rewriting, response rewriting, and both (retrieval recall is analysis-only and not attacker-observable), E-MIA holds an Accuracy of 0.988/0.987/0.988 with degradation of at most 0.003, whereas the baselines degrade sharply—S^2MIA loses up to 0.279 Accuracy under response rewriting (Table 2(a)).

Stealthiness against guardrail detectors. Table 2(b) confirms the stealthiness of E-MIA: by framing membership queries as ordinary educational "exams", it attains pass rates of 1.000, 0.997, and 1.000 on PIGuard, Tanaos-Guardrail-v1, and FT-Llama-Prompt-Guard-2, respectively, versus 0.210 for S^2MIA on the last detector. This benign pedagogical framing decouples the membership signal from the adversarial patterns that similarity-based and prompt-injection detectors flag.

Effect of the retrieval configurations. Holding the generator fixed, we jointly vary the retrieval budget $k = 1$ to 20 and swap the retriever across three embedding backends. As k grows, non-member documents add noise and baseline accuracy decays, whereas E-MIA remains stable up to $k = 20$ (Appendix Fig. 15). Baselines likewise fluctuate with embedding quality, while E-MIA stays near-optimal on every metric, including TPR at FPR 0.05, 0.01, and 0.005 (Appendix Fig. 16). Because each item targets specific document evidence, the inference is largely insensitive to both context and embedding backend, making E-MIA reliable across budgets and backends.

6 Conclusion

In this work, we reveal a practical membership leakage channel in black-box RAG systems: document-specific *hard evidence* can be reliably elicited through natural-looking QA prompts even when retrieval traces are hidden. To exploit this signal, we propose E-MIA, an exam-style document-level MIA framework that decomposes a target document into verifiable evidence, instantiates it into four objectively-gradable question types (FB/SC/MC/T/F), and aggregates black-box outcomes into a calibrated inference score for membership decisions. Extensive experiments across multiple datasets, retrievers, generators, and defense pipelines demonstrate that E-MIA achieves strong and stable separability, particularly under stringent low-FPR regimes, and remains robust to practical factors such as retrieval budget. These results highlight that RAG deployments may expose privacy risks beyond coarse similarity cues, underscoring the need for stronger mitigations against evidence-level leakage.

REFERENCES

Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. In Roberto Di Pietro, Karen Renaud, and Paolo Mori (eds.), *ICISSP*, pp. 474–485. SCITEPRESS, 2025. doi: 10.5220/0013108300003899. URL https://doi.org/10.5220/0013108300003899.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Atousa Arzanipour, Rouzbeh Behnia, Reza Ebrahimi, and Kaushik Dutta. Rag security and privacy: Formalizing the threat model and attack surface. In *2025 IEEE ICDMW*, pp. 1387–1394, 2025. doi: 10.1109/ICDMW69685.2025.00165.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *ICML*, pp. 2206–2240, 2022.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model. *arXiv preprint arXiv:2311.01149*, 2023.

Yuetian Chen, Kaiyuan Zhang, Yuntao Du, Edoardo Stoppa, Charles Fleming, Ashish Kundu, Bruno Ribeiro, and Ninghui Li. Membership inference attacks against fine-tuned diffusion language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oWKJursYpW.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *ICML*, pp. 1964–1974, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Sriram Dharwada, Himanshu Devrani, Jayant Haritsa, and Harish Doraiswamy. Query rewriting via llms, 2025. URL https://arxiv.org/abs/2502.12918.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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 *Australasian Conference on Information Security and Privacy*, pp. 147–166, 2025.

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 ACM SIGSAC*, pp. 4184–4198, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

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

Guoyao Li, Shahbaz Rezaei, and Xin Liu. User-level membership inference attack against metric embedding learning. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. URL https://openreview.net/forum?id=H__g7vTqIx9.

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. *arXiv preprint arXiv:2511.12043*, 2025a.

Hao Li, Xiaogeng Liu, Ning Zhang, and Chaowei Xiao. Piguard: Prompt injection guardrail via mitigating overdefense for free. In *Proceedings of ACL*, pp. 30420–30437, 2025b.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Effective passive membership inference attacks in federated learning against overparameterized models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QsCSLPP55Ku.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP*, pp. 1–5, 2025c.

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*, pp. 2894–2907, 2025.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of EMNLP*, pp. 5303–5315, 2023.

Aigerim Mansurova, Aiganym Mansurova, and Aliya Nugumanova. Qa-rag: Exploring llm reliance on external knowledge. *Big Data and Cognitive Computing*, 8(9):115, 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 *USENIX*, pp. 2369–2385, 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 ACM SIGSAC*, pp. 1245–1259, 2025.

Khanh Nguyen, Raouf Kerkouche, Mario Fritz, and Dimosthenis Karatzas. DocMIA: Document-level membership inference attacks against docVQA models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gNxvs5pUdu.

Aira Security. Ft-llama-prompt-guard-2: Fine-tuned prompt injection and jail break detector, 2024. URL https://huggingface.co/Aira-security/FT-Llama-Prompt-Guard-2.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE symposium on security and privacy*, pp. 3–18. IEEE, 2017.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

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. *arXiv preprint arXiv:2307.09288*, 2023.

Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness. In *Findings of the ACL*, pp. 13111–13128, 2024.

Guangshuo Wang, Jiajun He, Hao Li, Min Zhang, and Dengguo Feng. Rag-leaks: difficulty-calibrated membership inference attacks on retrieval-augmented generation. *Science China Information Sciences*, 68(6):160102, 2025a.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in NeurIPS*, 33:5776–5788, 2020.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI*, volume 39, pp. 25434–25442, 2025b.

Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Canary in a coalmine: Better membership inference with ensembled adversarial queries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b7SBTEBFnC.

YUTONG WU, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. You only query once: An efficient label-only membership inference attack. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7WsivwyHrS.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Peiru Yang, Jinhua Yin, Haoran Zheng, Xueying Bai, Huili Wang, Yufei Sun, Xintian Li, Songwei Pei, Yongfeng Huang, and Tao Qi. Mrm: Black-box membership inference attacks against multimodal rag systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34295–34303, 2026. doi: 10.1609/aaai.v40i40.40726.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models. *Association for Computing Machinery*, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A APPENDIX

Outline. This appendix is divided into nine sections:

- **Section A.1**: provides the full proofs of the expected membership score separation, the variance of the membership score, and the effect of exam length on score stability.
- **Section A.2**: provides the full related-work discussion that Section 2 condenses.
- **Section A.3**: details the datasets and preprocessing pipeline, the three baseline methods, and the evaluation metrics used in our E-MIA.
- **Section A.4**: demonstrates how our method extracts evidence from the target document and then generates the exam paper.
- **Section A.5**: details the three prompts used in our method during the implementation process.
- **Section A.6**: details the two-stage answer normalization and fault-tolerance strategy used for objective scoring.
- **Section A.7**: details how we conduct security detection experiments.
- **Section A.8**: details how we conduct rewriting defense strategy experiments.
- **Section A.9**: supplement the data with member score distribution plots for the NFCorpus and SCIDOCS datasets.
- **Section A.10**: reports generalization across retrieval configurations, languages, and exam generators.
- **Section A.11**: details the AI use statement.

A.1 PROOFS

This section provides concise proofs for the main formulations introduced in Section 4. We establish how the type-wise correctness probabilities determine the expected membership score, and how aggregating multiple questions reduces score variance.

A.1.1 PROOF OF MEMBERSHIP SCORE SEPARATION

Proposition 1 (Expected Membership Score)**.** *Let $\bar{S} = S/100 = \sum_{t\in\mathcal{T}} w_t \widehat{p}_t$, where $\widehat{p}_t$ is the empirical correctness rate for question type t. Then, for membership state $m \in \{0, 1\}$,*

$$\mathbb{E}[\bar{S} \mid M = m] = \sum_{t\in\mathcal{T}} w_t p_t^{(m)}. \tag{9}$$

Consequently, Eq. (7) holds, i.e.*,*

$$\Delta = \mathbb{E}[\bar{S} \mid M = 1] - \mathbb{E}[\bar{S} \mid M = 0] = \sum_{t\in\mathcal{T}} w_t \left(p_t^{(1)} - p_t^{(0)}\right). \tag{10}$$

Proof. Since $\widehat{p}_t$ is the average of the binary grading outcomes of type t,

$$\mathbb{E}[\widehat{p}_t \mid M = m] = p_t^{(m)}. \tag{11}$$

Therefore, by linearity of expectation,

$$\mathbb{E}[\bar{S} \mid M = m] = \sum_{t\in\mathcal{T}} w_t p_t^{(m)}. \tag{12}$$

Taking the difference between $M = 1$ and $M = 0$ yields Eq. (7). □

A.1.2 Proof of Score Variance

Proposition 2 (Variance of the Membership Score)**.** *Assume that question-level grading outcomes are conditionally independent given the membership status. Then Eq. (8) holds,* i.e.*,*

$$\mathrm{Var}(\bar{S} \mid M = m) = \sum_{t \in \mathcal{T}} \frac{w_t^2 p_t^{(m)}(1 - p_t^{(m)})}{N_t}. \tag{13}$$

Proof. For each question type t, $\widehat{p}_t$ is the average of N_t Bernoulli variables with success probability $p_t^{(m)}$. Hence,

$$\mathrm{Var}(\widehat{p}_t \mid M = m) = \frac{p_t^{(m)}(1 - p_t^{(m)})}{N_t}. \tag{14}$$

Because the type-wise estimates are conditionally independent,

$$\mathrm{Var}(\bar{S} \mid M = m) = \sum_{t \in \mathcal{T}} w_t^2 \,\mathrm{Var}(\widehat{p}_t \mid M = m), \tag{15}$$

which proves Eq. (8). □

A.1.3 Effect of Exam Length

Proposition 3 (Effect of Exam Length)**.** *For fixed weights and correctness probabilities, increasing the number of questions N_t for any question type t decreases the variance of the corresponding type-wise accuracy and therefore decreases the variance of the aggregated membership score.*

Proof. From Eq. (8),

$$\mathrm{Var}(\bar{S} \mid M = m) = \sum_{t \in \mathcal{T}} \frac{w_t^2 p_t^{(m)}(1 - p_t^{(m)})}{N_t}. \tag{16}$$

Each term is proportional to $1/N_t$. Thus, increasing N_t reduces the variance of the aggregated score. □

A.1.4 Implication for Membership Inference

The above results provide a simple interpretation of the E-MIA scoring mechanism. The expected separation between members and non-members is determined by the weighted type-level differences $p_t^{(1)} - p_t^{(0)}$, while the variance of the aggregated score decreases as more questions are evaluated. Therefore, under the evidence-discriminativeness premise $p_t^{(1)} > p_t^{(0)}$, E-MIA obtains a positive expected membership signal while question aggregation makes the resulting score more stable.

A.2 Related Work

Existing studies can be broadly categorized into two main lines: retrieval-augmented generation (RAG) frameworks and membership inference attacks in RAG systems.

A.2.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) mitigates knowledge staleness and hallucinations inherent to static parametric models by incorporating an external knowledge base at generation time, thereby providing large language models with retrievable evidence as contextual grounding Ji et al. (2023); Brown (2020); Touvron et al. (2023); Gao et al. (2024); Arzanipour et al. (2025). A typical RAG pipeline consists of a retriever and a generator: given a user query, the retriever returns the top-k relevant document passages from an external corpus, which are then concatenated with the query and fed into the generator to produce the final response. Because the model's output is explicitly conditioned on the retrieved content, the external document repository and its associated index constitute core assets of a RAG system Anil et al. (2023); Brown (2020); Touvron et al. (2023). This

dependency also introduces a new privacy exposure surface: even if the system does not directly disclose verbatim text, an adversary may still infer (solely via black-box interactions and output analysis) whether a particular document is included in the corpus or is retrievable by the system.

A.2.2 MEMBERSHIP INFERENCE IN RAG SYSTEMS (RAG-MIA)

Membership inference attacksHu et al. (2022); Choquette-Choo et al. (2021); He et al. (2025); Li et al. (2023; 2022); Chen et al. (2026) in the RAG setting (RAG-MIA) aim to determine whether a target document (or target sample) is present in the retrieval corpus/index of a RAG system, where the adversary typically has only black-box query access and can observe the system's outputs Liu et al. (2025); Naseh et al. (2025); Gao et al. (2025); Li et al. (2025a); Shokri et al. (2017); Feng et al. (2025); Nguyen et al. (2025). Existing approaches can be broadly categorized by the source of the membership signal into three classes: (i) explicit confirmation methods directly ask whether target content appears in the retrieved context (*e.g.*, requiring a Yes/No response). While such signals can be strong, the attack intent is often conspicuous and therefore prone to detection or refusal Anderson et al. (2025); (ii) similarity- or generation-quality scoring methods derive a membership score from semantic similarity between the output and a reference answer (or target text), perplexity, or related metrics. These methods are black-box friendly but can be confounded by model generalization in RAG settings Li et al. (2025c); Naseh et al. (2025); (iii) content probing methods amplify the discrepancy between members and non-members via masked completion of the target text, difficulty/perturbation calibration, or side-channel cues (*e.g.*, generation budgets). Although these techniques may improve separability, they often rely on stronger assumptions (*e.g.*, additional controllable interfaces, a reference system, or multi-round probing) and entail more complex execution pipelines Gao et al. (2025); Liu et al. (2025); Li et al. (2025a); Wang et al. (2025a); Wen et al. (2023).

Overall, practical RAG-MIA methods typically face a trade-off between robustness and feasibility. "Soft signals" based on semantic similarity or generation quality are susceptible to model generalization, leading to overlapping score distributions and unstable decision thresholds; conversely, enhancement strategies (*e.g.*, masked completion and calibrated perturbations) can improve discriminability but frequently introduce extra assumptions (*e.g.*, a reference model, controllable knobs, or multi-query overhead) and may be more likely to trigger defensive mechanisms Varshney et al. (2024). Departing from these paradigms, our work grounds membership inference in verifiable evidence intrinsic to the document by casting the task as an exam-style evaluation problem, using an aggregated exam score as the membership signal to achieve more stable statistical separation.

A.3 BASELINE AND EVALUATION METRICS

This section details the datasets and preprocessing pipeline, the three baseline methods, and the evaluation metrics used to assess the effectiveness of E-MIA.

A.3.1 DATASETS AND PREPROCESSING

We evaluate E-MIA on three corpora from the BEIR benchmark Thakur et al. (2021): NFCorpus[1], TREC-COVID[2], and SCIDOCS[3]. Table 3 summarizes their domain, scale, and average document length.

We apply a unified preprocessing pipeline to every corpus. We discard documents shorter than 50 tokens, which are too short to yield enough gradable evidence, and remove duplicates so that repeated content cannot inflate attack performance. Each preprocessed corpus is then split at random into 60% members, which are ingested into the RAG knowledge base, and 40% non-members, which are held out. For every dataset we draw a balanced evaluation set of 2,000 documents (1,000 members and 1,000 non-members). Because non-members are never ingested, any correct answer on their evidence must stem from parametric memorization or spurious correlation rather than from retrieval, which is what makes the balanced set a meaningful test of membership inference. Unless otherwise stated, all reported results are computed on this evaluation set.

[1] https://huggingface.co/datasets/BeIR/nfcorpus/

[2] https://huggingface.co/datasets/BeIR/trec-covid/

[3] https://huggingface.co/datasets/BeIR/scidocs/

Table 3: Characteristics of the studied datasets.

Datasets	Domain	Number	Avg Length	Example
NFCorpus	Bio-Medical	3.5K	1496.92	Recent studies have suggested that statins, an established drug group in the prevention of cardiovascular mortality...
TREC-COVID	Bio-Medical	23K	1356.24	This retrospective chart review describes the epidemiology and clinical features of 40 patients with culture-proven Mycoplasma pneumoniae...
SCIDOCS	Science	116K	1147.23	An evolutionary recurrent network which automates the design of recurrent neural/fuzzy networks using a new evolutionary learning algorithm...

A.3.2 Comparison Methods

We compare our proposed E-MIA against three state-of-the-art black-box membership inference attack methods designed for RAG systems: S^2MIA, MBA, and IA.

- **S^2MIA**: S^2MIA utilizes the first half of a target document as the query and computes the semantic similarity (*e.g.*, BLEU) between the RAG system's response and the original document as the membership score. It assumes that if a document exists in the database, the generated response will exhibit higher semantic overlap with the ground truth.
- **MBA**: MBA masks a fixed number of tokens within the target text and prompts the RAG system to complete them. Since RAG systems typically achieve higher completion accuracy on member samples, this accuracy is used as the membership signal.
- **IA**: Interrogative Attack (IA) employs an LLM to extract a summary from the target text and generate n Yes/No questions with corresponding ground-truth answers. Each question, combined with the summary, is submitted to the target RAG system. The membership signal is derived from the response accuracy across these queries, based on the intuition that member samples yield higher answer correctness.

A.3.3 Evaluation Metrics

To quantitatively evaluate the performance of the membership inference attack, we employ the following standard classification metrics:

- **Accuracy**: Accuracy measures the overall correctness of predictions in binary classification. Since the dataset is balanced between member and non-member samples, accuracy is equivalent to balanced accuracy in this setting.
- **AUC-ROC**: The Area Under the Receiver Operating Characteristic curve (ROC-AUC) measures the model's ability to distinguish between member and non-member samples across all classification thresholds. It reflects the probability that a randomly selected member sample is ranked higher than a non-member sample.
- **AUC-PR**: The Area Under the Precision-Recall curve summarizes the trade-off between precision and recall. This metric is particularly informative in security scenarios where both false positives and false negatives carry significant consequences.
- **TPR@FPR**: The True Positive Rate (TPR) at a fixed False Positive Rate (FPR) (*e.g.*, 0.05, 0.01, 0.005) evaluates the effectiveness of the attack under strict error constraints. This metric is especially important in practical attack settings, where maintaining a very low false positive rate is critical.

A.4 Document-to-Exam Transformation

We transform unstructured documents into structured exam queries based on five types of verifiable hard evidence, as illustrated in Fig. 8 and Fig. 9. To demonstrate the generality of the framework, we present both English and Chinese examples, showing that the proposed transformation is robust across languages and domains.



Figure 8: Document evidence extraction

A.4.1 Evidence Extraction

As shown in Fig. 8, documents inherently contain dense informational anchors that can be systematically extracted and categorized into five types: *Precise details*, *Proper names and technical terms*, *Definitional statements*, *Metadata cues*, and *Constraint relations*. These categories capture complementary aspects of document information, including numerical facts (*e.g.*, statistical values and time spans), domain-specific entities, conceptual definitions, structural or source-level identifiers, and logical dependencies (*e.g.*, conditional rules and criteria).

For example, in the English medical document, numerical statistics such as survival rates and time durations represent *Precise details*, while clinical adjustment conditions correspond to *Constraint relations*. In the Chinese advisory text, financial thresholds and decision rules exhibit similar patterns. Despite differences in language and writing style, both documents contain all five types of evidence, indicating that such informational structures are stable and language-agnostic. These extracted evidence units provide a reliable foundation for constructing discriminative membership signals.

A.4.2 Exam Generation

Based on the extracted evidence, we construct structured exam questions as illustrated in Fig. 9. Each type of evidence is mapped to a corresponding question format to ensure objective evaluation, following the same evidence-to-question mapping adopted in the main text. Specifically, *Proper names and technical terms*, together with local causal and *Constraint relations*, are converted into true/false items that test a single entity or the directionality of a logical link, whereas *Precise details*, *Definitional statements*, and *Metadata cues* are converted into fill-in-the-blank items that preserve exact numbers, concept boundaries, and structural markers. *Precise details* additionally supply at-

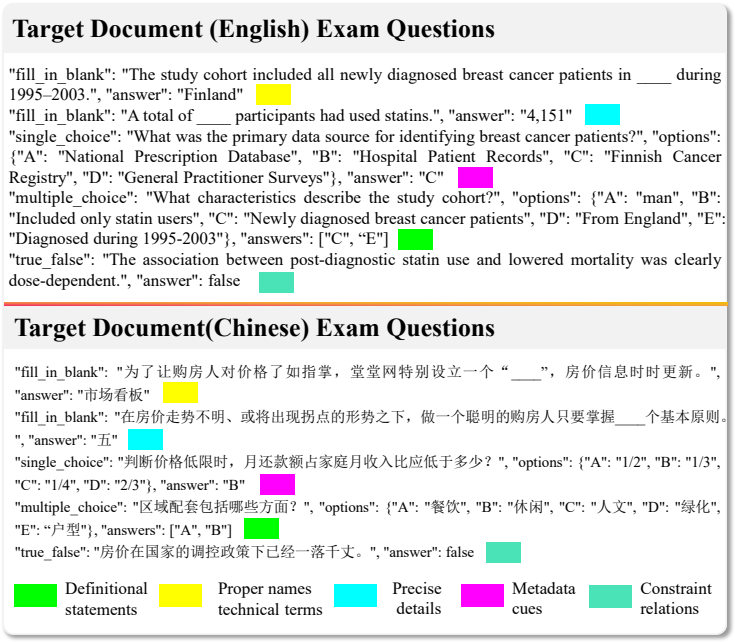


Figure 9: Exam example both English and Chinese language

tributes for multiple-choice items, in which several facts are aggregated into a single multi-attribute item.

This mapping strategy is consistently applied across languages. In both English and Chinese examples, the generated questions effectively probe the model's ability to recall factual information, understand logical relationships, and distinguish domain-specific entities. The resulting heterogeneous questions are then integrated into a unified exam set, enabling deterministic and interpretable scoring of membership signals while reducing sensitivity to linguistic variation.

A.5 Prompt Design for Membership Inference

We design three coordinated prompt modules to support the membership inference pipeline: (1) exam generation, (2) target RAG system configuration, and (3) inference-time query control. These prompts operate sequentially to transform documents into structured tests and enforce reliable model responses.

A.5.1 Exam Generator Prompt

The exam generator prompt (Fig. 10) instructs a high-capacity LLM to convert unstructured documents into structured exam questions. It enforces the generation of four complementary question types: fill-in-the-blanks (FB), multiple choice (MC), single choice (SC), and true/false (T/F), ensuring comprehensive probing of different evidence forms.

The prompt explicitly prioritizes five categories of verifiable hard evidence—*precise details*, *proper names and technical terms*, *definitional statements*, *metadata cues*, and *constraint relations*—as question anchors. To maximize discriminative power, each question type is subject to specific constraints: FB targets exact numerical values and technical terms; SC requires plausible but incorrect distractors; MC includes multiple correct answers to prevent guessing strategies; and T/F strictly adheres to document content. The output is formatted in a standardized JSON structure for automatic parsing and scoring.

A.5.2 RAG System Prompt

The RAG system prompt (Fig. 11) defines the operational behavior of the target model within a retrieval-augmented setting. It establishes a consistent system persona (*e.g.*, a helpful assistant) and specifies how external retrieved documents are incorporated into responses. This configura-

Exam Generator Prompt

You are a test question generation assistant.Your task is to generate a complete test paper strictly and completely based on the document provided below.

DOCUMENT: Title:{title} Text:{text}

Requirements:

1.Test Paper Structure

The test paper must contain the following four question types:

Fill in the Blanks (FB) Multiple Choice (MC) Single Choice (SC) True/False (T/F)

Each question type is designed to assess the candidate's different levels of understanding of the document. Questions can only be answered if the candidate accurately recalls the document's content.

·**FB**: Focuses on highly specific and detailed information in the document. Targets precise numerical values, dates, entity names, unique technical terms, identifiers, or metadata. Each blank must have only one clearly correct answer.

·**MC**: Each question must have four or five options, two of which are correct. Options should represent detailed attributes, characteristics, or conditions explicitly stated in the document.Avoid simple elimination strategies; a correct answer should require a precise understanding of the document's content.

·**SC**: Each question must have four or five options, only one of which is correct. Distractors should be plausible but clearly incorrect according to the document, and include one or two incorrect option similar to the correct one. Focus on testing the document's core facts, main definitions, or main arguments.

·**T/F**: Statements should test logical consistency and factual accuracy. Each statement must be explicitly supported or refuted by the document. Avoid vague or subjective statements.

2.Output Example:

"fill-in-blank":["question":"___directed Méditerranée (1963).", "answer":"Jean Daniel Pollet"],

"single-choice":["question":"Who directed Méditerranée (1963)? ", "options":{ "A" :"Jean-Daniel Pollet","B":"Stuart Rosenberg", "C" :"François Truffaut", "D" :"Alan Resnais"},"answer": "A"],

"multiple-choice":["question":"Which films are French?","options":{"A":"Move", "B":"Méditerranée","C":"Breathless","D":"The 300 Blows", "E": "Last Year at Maritin "},"answers": {“B","C"}],

"true/false": ["question": "Move (1970) was directed by an American director.", "answer": true]

Figure 10: Exam Generator Prompt

tion reflects realistic deployment conditions, ensuring that membership inference is evaluated under practical RAG usage.

By standardizing how retrieved knowledge interacts with model reasoning, this prompt enables controlled measurement of how database membership influences response accuracy. It therefore serves as the environmental baseline for evaluating the impact of stored documents on model outputs.

RAG System Prompt

You are a helpful assistant. Below are questions asked retrieved by users and related information from the database. Please answer the questions based on this information and your own reasoning.

Contexts: **[context]**

Query: **[question]**

Answer:

Figure 11: RAG System Prompt

A.5.3 LLM Query Prompt

The LLM Query Prompt (Fig. 12) governs the inference phase by enforcing a strict “exam-style” answering protocol. It removes conversational flexibility and constrains the model to produce precise, deterministic outputs aligned with the generated questions.

The prompt prohibits external tools, browsing, and speculative reasoning, ensuring that responses rely solely on internal knowledge or retrieved context. It further enforces strict formatting rules, including exact answer matching, prohibition of paraphrasing, and set-based selection constraints for multi-choice tasks. These restrictions eliminate semantic ambiguity and prevent the model from leveraging linguistic generalization to bypass factual queries.

As a result, this prompt maximizes the fidelity of membership signals and provides a reliable interface between the generated exam and the final scoring process.

LLM Query Prompt

We are now in an exam setting. Please complete the test according to the exam requirements.

INSTRUCTIONS:
Answer all questions strictly based on your knowledge. Do NOT use external tools, browsing, or assumptions beyond what you know. If you are uncertain about an answer, provide your best possible guess. Do NOT leave any question unanswered.

ANSWERING RULES:
Fill-in-the-Blank (FB):
Provide the exact missing content. Use the same wording, terminology, and formatting as in the original document. Do NOT paraphrase or explain.

Multiple-Choice (MC):
Select ALL options that you believe are correct. Do NOT select options unless you are confident they are supported.

Single-Choice (SC):
Select exactly ONE option. Choose the option that best matches the document content.

True/False (T/F):
Answer each statement with either "True" or "False" only.

Figure 12: LLM Query Prompt

A.6 Answer Normalization and Fault-Tolerance Strategy

To ensure reliable evaluation, we adopt a two-stage strategy that combines prompt-level constraints with post-hoc normalization. First, during inference, we enforce strict formatting requirements through the query prompt, instructing the model to produce canonical answers (*e.g.*, single-letter outputs for multiple-choice questions). These constraints guide the model toward structured and machine-parseable responses, reducing variability at the source.

Despite these constraints, language models may still generate loosely formatted or non-compliant outputs due to generation variability or instruction deviation. To address this, we introduce a fault-tolerant normalization layer that recovers intended answers without compromising scoring objectivity. This stage applies rule-based extraction, pattern matching, and semantic-preserving transformations. Specifically, for single- and multiple-choice questions, regular expressions identify the first valid option pattern (*e.g.*, A–D, a–d, 1–4), ignoring any trailing explanations. A mapping table further normalizes common paraphrases (*e.g.*, "the second option" $\rightarrow$ "B", "both A and C" $\rightarrow$ "AC"), while numerical ranges (*e.g.*, "42–45") are expanded into discrete sets. For fill-in-the-blank and true/false tasks, we apply text normalization, including case unification, punctuation removal, abbreviation expansion (*e.g.*, "Feb." $\rightarrow$ "February"), and standardized date/number formatting (*e.g.*, "3/15/2023" $\rightarrow$ "2023-03-15").

These mechanisms ensure that evaluation results are invariant to superficial formatting differences, providing a faithful measurement of the model's factual recall rather than its output style.

A.7 Security Detection Setup

We evaluate the stealthiness and prompt-evasion capability of E-MIA by testing whether attack queries can bypass existing guardrail systems without being flagged.

We consider three representative detection models: **PIGuard**, **Tanaos-Guardrail-v1**, and **FT-Llama-Prompt-Guard-2 (Prompt-Guard)**. PIGuard incorporates a Mitigating Over-defense for

Table 4: Example exact queries generated by each of the baseline attacks (S^2MIA, MBA, IA) and our attack (E-MIA) for the same target document

Attack	Query
Target Document	Epilepsy or seizure disorder is one of the most common neurological diseases in humans. Although genetic mutations in ion channels and receptors and some other risk factors such as brain injury are linked to epileptogenesis, the underlying cause for the majority of epilepsy cases remains unknown. Gene-environment interactions are thought to play a critical role in the etiology of epilepsy. Exposure to environmental chemicals is an important risk factor. Methylmercury (MeHg) is a prominent environmental neurotoxicant, which targets primarily the central nervous system (CNS).
S^2MIA	Given the [Query]: "Epilepsy or seizure disorder is one of the most common neurological diseases in humans. Although genetic mutations in ion channels and receptors and some other risk factors such as brain injury are linked to epileptogenesis, the underlying cause for the majority of epilepsy cases remains unknown". Do not include any introductory or explanatory text, use the following format for output: [Response]: 'Provide a concise response directly addressing the [Query] by using the most relevant and matching text in the prompt.'
MBA	Predict the answers of the masks in the given texts: Epilepsy or [MASK_1] disorder is one of the most common neurological diseases in humans. Although genetic mutations in ion channels and receptors and some other risk factors such as brain injury are linked to [MASK_2], the underlying cause for the majority of epilepsy cases remains unknown. [MASK_3]-environment interactions are thought to play a critical role in the etiology of epilepsy. Exposure to environmental chemicals is an important risk factor. [MASK_4] (MeHg) is a prominent environmental neurotoxicant, which targets primarily the [MASK_5] nervous system (CNS). You should strictly follow the contexts below to answer, and provide your answers in the specified format:"[Mask_i]: answer_i" for each mask:
IA	I want to ask about epilepsy as a common neurological disease, its potential causes including genetic and environmental factors, and the role of methylmercury. Is epilepsy primarily linked to genetic mutations in ion channels and receptors? Please answer with "Yes," "No," or "I don't know."
E-MIA(Ours)	We are now in an exam setting. Please complete the test according to the exam requirements. "Question": "Epilepsy or ____ is one of the most common neurological diseases in humans.", Provide the exact missing content. Use the same wording, terminology, and formatting as in the original document. Do NOT paraphrase or explain. "Question": "What is one of the most common neurological diseases in humans", "Options": {"A": "Epilepsy or seizure disorder", "B": "Alzheimer's disease", "C": "Parkinson's disease", "D": "Multiple sclerosis"}, You MUST choose exactly one option. Answer ONLY with the letter of the option. "Question": "What is stated about the cause of epilepsy for the majority of cases?", "options": {"A": "The underlying cause remains unknown", "B": "Genetic mutations are the sole cause", "C": "Brain injury is the most common cause", "D": "Gene-environment interactions are thought to be critical", "E": "It is always due to environmental chemicals"}, You MUST choose two options. Answer ONLY with letters of all options, separated by commas. "Question": "Epilepsy is one of the most common neurological disorders in humans, is true or false?", Answer ONLY with True or False.

Free (MOF) strategy to reduce bias toward trigger words. Tanaos-Guardrail-v1 serves as a comprehensive safety filtering system for harmful interaction detection. Prompt-Guard is a LoRA fine-tuned version of Llama-Prompt-Guard-2, optimized for prompt injection and jailbreak detection. We use the public checkpoints `leolee99/PIGuard`, `tanaos/tanaos-guardrail-v1`, and `Aira-security/FT-Llama-Prompt-Guard-2`, respectively.

For evaluation, we simulate realistic attack scenarios by generating 1,000 query instances for each method, including S^2MIA, MBA, IA, and our proposed E-MIA (see Table 4). These queries are fed into each guardrail model to assess whether they are flagged or allowed. We adopt **Pass Rate** as the evaluation metric, defined as the proportion of queries that successfully bypass the guardrails. A higher pass rate indicates stronger stealthiness and better evasion capability of the attack.

A.8 Rewriting Defense Setup

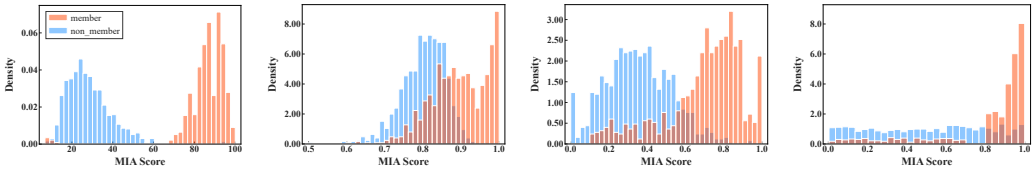
We evaluate the robustness of membership inference attacks under semantic perturbations introduced by rewriting-based defense strategies.

We consider three rewriting configurations: **Query Rewriting**, **Response Rewriting**, and **Both Rewriting**. Query Rewriting paraphrases the attacker's input (*e.g.*, using GPT-4o) to alter its surface form while preserving semantics, aiming to disrupt prompt-dependent attacks. Response Rewriting applies similar transformations to the RAG system output, weakening observable membership signals. Both Rewriting represents the strongest defense setting, where both input queries and responses are perturbed.

For each attack method (S^2MIA, MBA, IA, and E-MIA), we generate 1,000 query instances and process them through the RAG pipeline under each rewriting configuration. We evaluate performance using two metrics: **Retrieval Recall**, which measures whether the correct member document is retrieved, and **Accuracy**, which reflects the attack's ability to correctly infer membership under semantic perturbations.

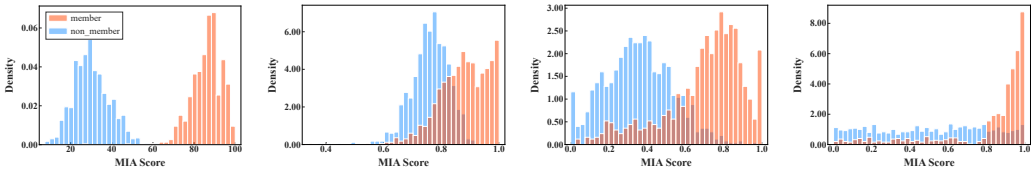
A.9 Membership Score Distributions Across Datasets

This section presents additional distribution analyses on NFCorpus and SCIDOCS (Fig. 13 and Fig. 14) to complement the primary results on TREC-COVID. These experiments aim to evalu-



(a) E-MIA(ours) (b) S^{2}MIA (c) MBA (d) IA

Figure 13: Membership and non-membership score distribution (NFCorpus)



(a) E-MIA(ours) (b) S^{2}MIA (c) MBA (d) IA

Figure 14: Membership and non-membership score distribution (SCIDOCS)

ate whether the separation between member and non-member samples remains consistent across datasets with different domains and statistical properties.

The distribution plots show the frequency of accuracy scores obtained during the automated examination phase. For member samples, the scores are consistently skewed toward the high-accuracy region, indicating that the model can reliably reproduce the extracted hard evidence and logical constraints. In contrast, non-member samples exhibit lower and more dispersed accuracy patterns, reflecting reliance on zero-shot generalization rather than memorized content. This results in a clear separation between the two groups, often forming a bimodal distribution.

Such separation directly explains the strong AUC-ROC performance of E-MIA, as it reduces overlap between member and non-member score distributions and mitigates false positives commonly observed in semantic-similarity-based attacks. Importantly, this pattern remains stable across NFCorpus and SCIDOCS despite differences in terminology density and document structure. This consistency suggests that memorization of hard evidence is a persistent and domain-agnostic phenomenon, supporting the robustness of our approach for auditing privacy risks in diverse RAG systems.

A.10 Generalization Analysis

We evaluate E-MIA along four axes that stress different aspects of deployment: retrieval budget, retriever, corpus language, and exam generator.

Retrieval budget. Increasing k injects more non-member documents into the retrieved context, and the baselines lose accuracy accordingly, whereas E-MIA remains near-optimal up to $k = 20$. This divergence reflects a fundamental difference in how the two paradigms extract membership signals. Baselines typically rely on output-level similarity, perplexity, or confirmation cues computed over the entire retrieved context, so a larger k dilutes the signal-to-noise ratio: relevant evidence is buried among irrelevant documents, and the model's output becomes increasingly influenced by non-member content. As k grows, this dilution compounds, causing baseline accuracy to degrade steadily (Fig. 15).

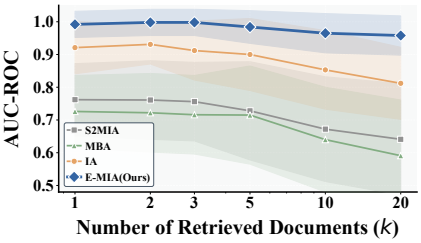
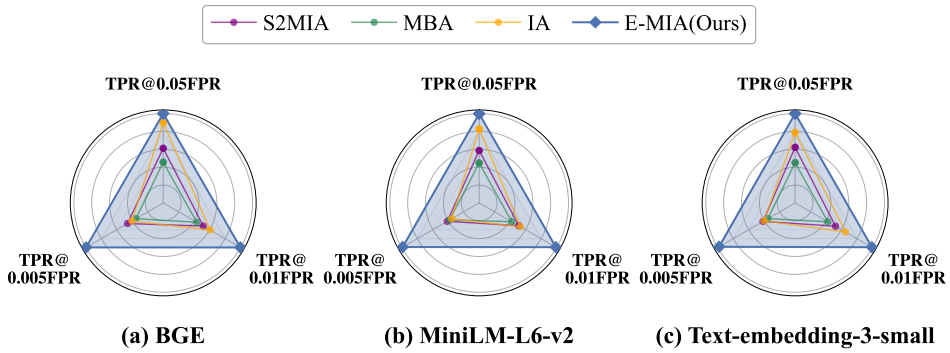


Figure 15: Retrieval budget

E-MIA, by contrast, is designed around document-specific evidence rather than holistic context. Each exam item probes one precise fact or relation drawn from the target document, and grading is performed against an objective answer key independent of the surrounding passages. Consequently, the presence of additional non-member documents has little effect on whether the model answers a given item correctly: what matters is whether the target evidence was memorized during training, not whether it is salient in the current retrieval context. Because each exam item targets one

specific piece of document evidence, the inference is largely insensitive to the surrounding context. This property makes E-MIA robust to retrieval budget variation, and explains why its performance remains near-optimal even when k is increased to 20.

Various Retrievers. Holding the generator fixed, we swap the retriever across BGE, MiniLM-L6-v2, and Text-embedding-3-small. While the baselines fluctuate with the quality of the embedding space, E-MIA stays near-optimal on every metric, including TPR at FPR levels 0.05, 0.01, and 0.005 (Fig. 16). Because each item targets document-specific evidence, effectiveness is largely independent of the initial document ranking, which makes E-MIA reliable across retrieval backends.



(a) BGE **(b) MiniLM-L6-v2** **(c) Text-embedding-3-small**

Figure 16: Performance of E-MIA across BGE, MiniLM-L6-v2, and Text-embedding-3-small retrievers.

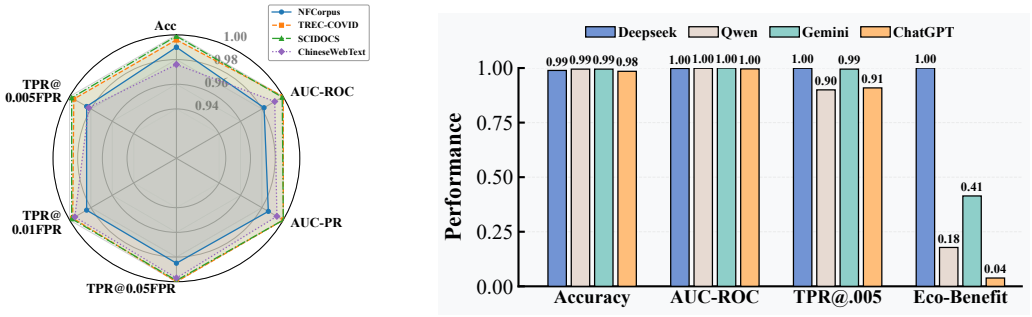
Performance on Chinese Datasets. Extending the evaluation to the Chinese corpus ChineseWebText Chen et al. (2023), E-MIA reaches near-perfect Accuracy, AUC-ROC, and TPR at all FPR levels with no language-specific degradation (Fig. 17(a)). This result is notable because Chinese and English differ substantially in tokenization, morphology, and syntactic structure, yet the exam-based membership signal transfers almost perfectly across the two languages. The close agreement between the English and Chinese results indicates that the membership signal stems from the model's intrinsic memorization rather than from surface-level linguistic features, so the exam remains a language-agnostic auditing tool.

The underlying reason is that E-MIA does not rely on any language-specific cue such as perplexity patterns, fluency, or stylistic consistency. Instead, each exam item is constructed from document-specific evidence and graded against an objective answer key, so the resulting score depends only on whether the model has memorized the target content during training. Since memorization is a property of the training data rather than of the surface language, the same evidence-to-question mapping yields comparable signal strength across languages.

This language-agnostic property is particularly valuable for auditing multilingual foundation models, where training corpora are often mixed and language-specific heuristics may not generalize.

Effect of the Exam Generator Model. To verify that E-MIA does not depend on a particular generative architecture, we also generate exams with Gemini-2.0-Flash Comanici et al. (2025), Qwen-Max Yang et al. (2025), and GPT-5-mini Singh et al. (2025). These generators differ substantially in parameter scale, training data, and instruction-following behavior, so any architecture-specific bias in exam construction would manifest as a visible gap in downstream auditing performance. Accuracy stays within 0.985–0.995, AUC-ROC within 0.996–1.000, and TPR@FPR$=0.005$ above 0.900 for every generator (Fig. 17(b)). The narrow spread across all three metrics indicates that the membership signal is governed by the underlying evidence-to-question mapping rather than by the stylistic or structural quirks of any single generator.

This robustness stems from the fact that exam items are constrained by an objective scoring rule and a fixed answer key, which limits the space of valid questions a generator can produce. As a result, different generators converge on functionally equivalent exams even when their surface wording differs, and the resulting scores remain comparable across architectures. We nevertheless adopt DeepSeek-R1 as our default generator: with Eco $=$ $\text{Cost}_{\min}/\text{Cost}_{\text{model}}$, it is the cheapest of the four at $\$0.0063$ per target document while matching the others in effectiveness. This cost advantage matters in practice, since auditing a large corpus requires generating exams for many



(a) Cross-lingual performance (b) Performance across exam generators

Figure 17: Generalization of E-MIA across languages and exam-generator models.

target documents, and generator cost scales linearly with corpus size while providing no additional signal once effectiveness saturates.

A.11 AI USE STATEMENT

Generative AI tools were used solely to assist with code development and debugging. All AI-assisted code was reviewed and tested by the authors. The authors take full responsibility for the paper's content, claims, experimental results, and artifacts.