HoneyMark: Protecting Data Integrity in RAG via Synthetic Trap Documents and Post-hoc Watermarking

Zhuo Wang ✉, Guixuan Zhang † ✉, Jiayang Yu ✉, Shuwu Zhang ✉

Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract

Data owners lack effective methods to detect unauthorized incorporation of their content into Retrieval-Augmented Generation (RAG) systems. Existing approaches like WARD (Watermark-based RAG Dataset inference) compromise data integrity by modifying original documents with watermarks prone to signal loss. We propose HoneyMark, a trap document-based approach that achieves dataset membership inference without altering original data. HoneyMark generates synthetic documents with POSTMARK post-hoc watermarks and inserts them as traps into datasets. When queried about synthetic content, RAG systems containing these documents reveal watermark vocabulary in responses, providing statistical evidence of unauthorized usage. Our four-dimensional generation framework creates trap documents semantically consistent with original data while maintaining content uniqueness. Experimental results demonstrate perfect discrimination (AUC-ROC = 1.0) with only 16 queries and 100% TPR at 1% FPR with 20 queries, significantly outperforming existing methods while preserving data integrity.

Keywords: Retrieval-Augmented Generation, Dataset membership inference, Post-hoc watermarking, Trap documents, Data integrity

1. Introduction

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by incorporating retrieval mechanisms that enable dynamic access to external corpora[1], effectively mitigating issues such as model hallucination, outdated knowledge, and untraceable generation results[2]. This approach demonstrates significant advantages in tasks including open-domain question answering, information retrieval, and knowledge management[3][4]. However, to ensure system timeliness and reliability, RAG providers must continuously update their external corpora, which may raise concerns regarding data compliance and copyright protection, particularly when publicly accessible documents are used without proper authorization[5]. Notably, data owners typically lack effective methods to detect whether their content has been incorporated without authorization into third-party RAG systems' retrieval corpora, making intellectual property status difficult to trace and verify[6]. Therefore, reliably verifying whether a specific dataset is being used by a target RAG system has become a critical challenge that urgently needs to be addressed.

†Corresponding author: Guixuan Zhang (Email: guixuan.zhang@bupt.edu.cn)

The core difficulty in verifying whether a dataset has been incorporated into a Retrieval-Augmented Generation (RAG) system lies in the fact that both the corpus and retrieval mechanisms typically operate as invisible black boxes[7]. Data owners can only make indirect inferences by inputting queries and analyzing the generated responses[8], which not only significantly increases the complexity of the verification problem but also renders existing compliance and copyright protection mechanisms ineffective. Previous research has attempted to address this issue through the WARD framework[9], which rewrites each document in the dataset using watermarked large language models, thereby implicitly embedding invisible red-green watermark signals in the text and achieving statistical detection[10] of watermark signals through analysis of RAG-LLM system outputs. However, direct rewriting of original documents may introduce factual deviations, thereby undermining the integrity and reliability of documents[11] in terms of knowledge representation and information transmission. Additionally, the red-green watermarking mechanism employed by WARD intervenes in token probability distributions during generation to embed implicit signals[12], potentially compromising text naturalness and linguistic fluency. More critically, in typical retrieval-augmented generation workflows, the local generative distribution features relied upon by red-green watermarks are prone to dilution or elimination[13], making watermark signals difficult to stably preserve in actual retrieval and generation pipelines, thus reducing detection robustness and efficiency. These limitations highlight unresolved issues in RAG dataset verification: how to maintain watermark signal detectability and robustness in complex retrieval and generation pipelines without compromising the factual content of documents.

To address these challenges, this paper proposes HoneyMark, a trap document-based RAG dataset membership inference strategy[11][8]. The core idea of this method is to achieve provable detection of unauthorized data usage in RAG systems by inserting carefully constructed synthetic documents carrying POSTMARK post-hoc watermarks[14] as traps, without modifying the original dataset content. Compared to existing methods, HoneyMark offers three significant advantages: First, it completely preserves the integrity of the original dataset, avoiding factual errors caused by document rewriting[15]; Second, POSTMARK's semantics-based watermarking mechanism demonstrates stronger stability[16] in complex RAG processing workflows; Third, the uniqueness of synthetic entities ensures high-precision targeted detection, effectively improving retrieval efficiency.

The main contributions of this paper include: (1) proposing a RAG dataset copyright verification method that requires no modification of original data; (2) designing a systematic four-dimensional synthetic document generation framework; (3) establishing statistical detection based on POSTMARK watermarks; (4) comprehensively validating the method's advantages in accuracy and detection efficiency through extensive experiments. Experimental results demonstrate that HoneyMark achieves reliable detection performance while maintaining data integrity.

2. Related Work

2.1. RAG Dataset Inference and Copyright Protections

The detection of unauthorized dataset usage in RAG systems represents an emerging yet critical research problem. Jovanović et al. [9] first formally defined the RAG Dataset Inference (RAG-DI) task, which aims to determine whether a data owner's dataset is included in a RAG system's retrieval corpus through black-box queries. Existing RAG-DI methods can be broadly categorized into two types. Li et al. [17] and Anderson et al.[8] proposed aggregation-based

inference approaches that essentially extend document-level membership inference attacks to the dataset level, making dataset-level judgments by aggregating inference results from multiple documents[17]. However, these methods perform poorly when confronted with the factual redundancy commonly present in real-world scenarios, as semantically similar documents lead to significant false positives[18]. WARD introduced a proactive protection approach that achieves copyright protection by embedding red-green watermarks into the original dataset. While this method provides theoretical guarantees, it faces several challenges in practical applications: modifications to original data compromise data integrity, and red-green watermarks rely on subtle changes in token generation probabilities. These implicit signals are highly susceptible to dilution or elimination after passing through multiple stages of RAG systems[19], including retrieval, rewriting, and generation, thereby limiting their practical effectiveness. More importantly, not only compromise the integrity and utility of the original data but also impose additional quality control costs on data owners. The fundamental limitation of current research lies in its failure to balance the contradiction between detection effectiveness and data integrity protection. Existing methods either suffer from insufficient detection accuracy or achieve performance at the cost of data quality.

2.2. Evolution of LLM Watermarking Technologies

The multi-stage text processing characteristics of RAG systems pose unique challenges for LLM watermarking technologies[20][21]. Unlike direct detection of LLM-generated text, RAG-DI requires watermark signals to remain detectable through operations such as retrieval, truncation, and reorganization[22][3], resulting in significantly different applicability of various watermarking schemes in RAG scenarios.

2.2.1. Generation-Time Watermarks (Red-Green Approach)

Generation-time watermarks, exemplified by the red-green watermarking (commonly referred to as KGW, named after Kirchenbauer, Geiping, Wen, et al.) proposed by Kirchenbauer et al.[23][24], embed signals by manipulating token selection probabilities during the decoding process. Specifically, this method divides the vocabulary into "green" (encouraged) and "red" (suppressed) subsets at each generation step[25] based on the preceding context and a secret key, enhancing the probability of green tokens during sampling. Detection is performed by calculating the z-score of green token proportions to determine whether text contains watermarks. However, such methods face fundamental challenges in RAG scenarios. First, the signal of red-green watermarks manifests as a subtle probabilistic bias that is highly susceptible to degradation during the multi-stage processing inherent in RAG systems[26][27][19][28]. When watermarked documents undergo retrieval, truncation, and reorganization before being input to generative models, the original token sequence structure is disrupted, and probabilistic bias information is consequently diminished.

2.2.2. Advantages of Post-hoc Watermarks

POSTMARK represents an emerging class of post-hoc watermarking technologies that achieve watermark embedding by inserting semantically relevant specific vocabulary into text after generation completion[14]. This method first selects a set of related vocabulary based on the text's semantic embeddings, then utilizes instruction-following models to naturally integrate these words into the original text. Compared to red-green watermarks, POSTMARK's robustness advantages stem from differences in signal carriers[14][16]. Red-green watermarks rely on

token selection probability biases at specific positions, and signals are lost when text structure changes. In contrast, POSTMARK's watermark signals manifest as concrete semantically relevant vocabulary. Even when these words change position during paraphrasing processes, due to their semantic association with the original text's theme, they still appear in rewritten text as synonyms or related concepts, thus maintaining detection feasibility. Additionally, the modular design ensures that this method does not require access to model logits and can be applied as an independent component to existing text[14], providing convenience for practical deployment.

2.3. Honeypot and Trap Techniques

In the field of digital copyright protection, trap techniques serve as a class of passive monitoring methods that have been widely applied to detect unauthorized data usage. Given the impact of traditional watermarking modification schemes on data integrity in RAG-DI scenarios and the limitations in precision of aggregation-based inference methods, trap techniques provide a new technical pathway for achieving the dual objectives of protecting original data while enabling reliable detection[6].

2.3.1. Honeypot Techniques in Data Leak Detection

In the domain of data leak detection, honeypot technology monitors abnormal access behavior by inserting decoy records into sensitive datasets[29]. Unlike traditional access control mechanisms, honeypot techniques employ a "trapping" strategy: normal users will not encounter these records, while malicious accessors inevitably trigger traps when obtaining data in bulk. Bowen et al.[30]. applied this concept to cloud storage environments, detecting insider threats by mixing honeypot files into user files. When these files are accessed abnormally, the system can immediately identify potential data breach behaviors. Similarly, in machine learning model protection, researchers have proposed inserting specially crafted samples into training data to detect unauthorized model usage[31].

2.3.2. Adaptation Challenges in RAG Scenarios

Applying trap techniques to RAG-DI scenarios faces unique challenges. Traditional trap techniques primarily target direct copying or access of data[32][33][34], while RAG systems utilize data through semantic retrieval and generative rewriting, making static trap information[35][36] based on simple keywords or fixed patterns potentially lost or deformed during processing. Furthermore, RAG systems' retrieval mechanisms are based on semantic similarity, requiring trap information to maintain semantic consistency with normal data[37][38] at the semantic level; otherwise, it cannot be effectively retrieved and triggered[39]. POSTMARK watermarks construct watermark signals by embedding semantically relevant vocabulary[14][40][41]. When synthetic documents are processed by RAG systems[42][43][44], even if original vocabulary changes, words semantically related to the document's theme will still appear in various forms[45] in the output, thus maintaining the possibility of statistical detection[46].

3. Method

This section provides a detailed introduction to the HoneyMark method proposed in this paper. Section 3.1 first introduces the overall architecture, followed by Section 3.2 which elaborates on the synthetic document generation framework, and finally Section 3.4 describes the statistics-based detection mechanism.

3.1. HoneyMark Architecture

The core of the HoneyMark method is to transform the RAG dataset inference problem into a trap detection problem.

Data Owner: Before publishing their dataset D_{do}, the data owner uses the framework described in Section 3.2 to generate a set of watermarked synthetic documents $D_{synthetic}$. Subsequently, the original dataset is merged with the synthetic documents to form an enhanced dataset $D_{enhanced} = D_{do} \cup D_{synthetic}$, which is then published.

Auditor: When suspecting that a RAG system M^* has used D_{do} without authorization, the auditor generates a series of queries Q targeting the synthetic facts in $D_{synthetic}$. By sending queries to M^* and analyzing its responses R, the auditor detects whether they contain POSTMARK watermark signals[47][25], thereby making an infringement determination.

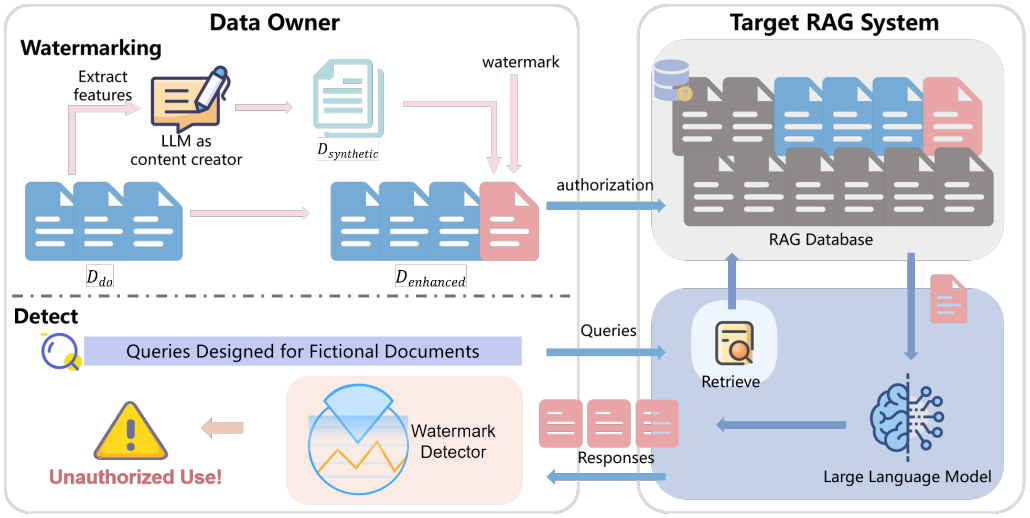


Figure 1: Overview of the HoneyMark framework. The data owner generates watermarked synthetic documents and merges them with the original dataset. Auditors query suspected RAG systems about synthetic entities; watermark vocabulary in responses enables statistical detection. Original data remains unmodified.

As show in Figure 1, the core advantage of this architecture lies in achieving "zero modification": the original dataset D_{do} remains completely unchanged, with all detection logic encapsulated within the expendable synthetic documents $D_{synthetic}$.

3.2. Synthetic Document Generation Framework

To ensure that trap documents can be effectively retrieved by RAG systems while maintaining the integrity of the original dataset, this paper proposes a synthetic document generation strategy based on four core dimensions: by systematically controlling the dimensions of topic, entity, time, and space, we generate trap documents that are semantically relevant to the original dataset but have completely synthetic content. As shown in Figure 2, the synthetic document generation method S comprises five sequential stages: 1) document feature extraction, 2)

conceptual entity construction, 3) spatiotemporal background generation, 4) content synthesis, and 5) watermark integration. The output of each stage provides constraints for subsequent stages, ensuring controllability and consistency in the generation process.

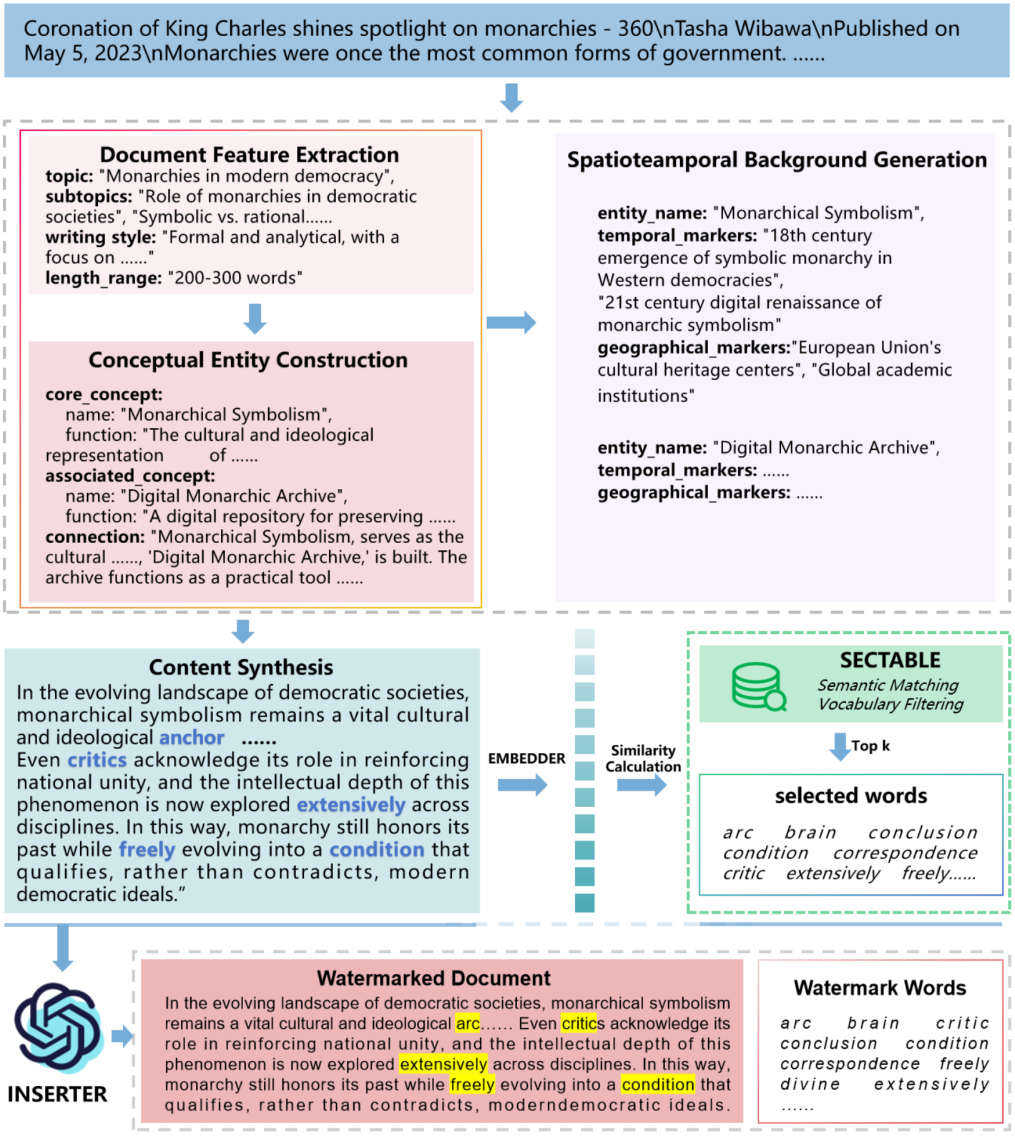


Figure 2: Four-Dimensional Synthetic Document Generation and Watermarking Pipeline.Five-stage pipeline: (1) Feature extraction, (2) Entity synthesis, (3) Spatiotemporal grounding, (4) Content synthesis, (5) POSTMARK watermarking. Example: "Monarchical Symbolism" document with highlighted watermark words.

Document Feature Extraction To ensure that synthetic documents maintain feature consistency with the original dataset D_p, the generation process first randomly samples a document $d \in D_p$,

then uses LLM M to extract its key attribute set $A = \{A_1, A_2, A_3, A_4\}$:

$$A \leftarrow M(\text{attr_extract_func}(d)). \quad (1)$$

Specifically, four types of attributes are extracted: A_1 = topic domain, A_2 = subtopic list, A_3 = writing style, A_4 = length range. These attributes constitute the fundamental constraints for subsequent generation processes, ensuring that synthetic documents align with the original dataset at the surface feature level.

Conceptual Entity Construction To establish detectable copyright markers, it is necessary to generate conceptual entities with unique characteristics as query targets. Based on the extracted attributes A, we use LLM to generate two types of synthetic entities: core concepts E (primary research objects) and associated concepts $\tilde{E}$ (supporting elements), while defining relationships R between them (representing business or technical connections between two concepts):

$$\{E, \tilde{E}, R\} \leftarrow M(\text{entity_synth_func}(A)). \quad (2)$$

Entity generation follows domain naming conventions and professional expression habits, but with completely synthetic content, ensuring no overlap with real-world entities and providing unique detection targets for subsequent targeted queries.

Spatiotemporal Background Generation To enhance the realism of synthetic documents and establish internal logical consistency, each entity is assigned a spatiotemporal marker set $C = \{C_t, C_g\}$, where C_t represents temporal markers and C_g represents geographical markers:

$$C \leftarrow M(\text{context_synth_func}(E, \tilde{E}, A)). \quad (3)$$

Temporal markers follow reasonable chronological order of technological development, while geographical markers use synthetic but geographically logical location information, ensuring internal consistency of the spatiotemporal background.

Content Synthesis The structured outputs from the preceding stages are integrated into a complete natural language document. This process requires entity names to appear naturally in the document, spatiotemporal information to be seamlessly embedded, and the overall narrative logic to be coherent:

$$d_s \leftarrow M(\text{article_synth_func}(E, \tilde{E}, C, A)). \quad (4)$$

Watermark Integration The generated synthetic document d_s undergoes two-stage watermark processing through the POSTMARK algorithm: first, relevant vocabulary is selected from a pre-built vocabulary based on the document's semantic vectors, then these words are naturally inserted into the document through LLM rewriting[14]:

$$d_w \leftarrow \text{POSTMARK_insert}(\text{POSTMARK_select}(d_s), r). \quad (5)$$

where r is the insertion ratio parameter, controlling the balance between watermark strength and text quality.

Compared to existing methods, the main advantage of this framework lies in achieving decoupling between content synthesis and feature consistency: surface consistency is ensured through

feature extraction, content uniqueness is guaranteed through entity synthesis, and logical credibility is enhanced through spatiotemporal settings. This design enables synthetic documents to be indistinguishable from real documents at a professional level while maintaining complete content independence.

3.3. POSTMARK Watermarking

3.3.1. Application of POSTMARK Technology in Synthetic Documents

In the HoneyMark framework, POSTMARK watermarking technology is specifically used to watermark system-generated synthetic documents rather than modifying the original dataset. This design ensures that the integrity and authenticity of the original dataset are completely preserved while embedding detectable watermark signals in the synthetic "trap" documents. POSTMARK employs a post-hoc processing approach based on semantic embeddings, comprising three core components: the semantic embedder (EMBEDDER), the secret vocabulary table (SECTABLE), and the text processor (INSERTER)[14]. The key advantage of this technology lies in its ability to achieve watermark embedding without requiring access to the internal parameters of language models, solely through subsequent processing of generated text.

3.3.2. Semantic-based Watermark Vocabulary Determination

The watermark vocabulary selection process is a critical component that determines both the robustness and semantic coherence of the watermark. This process employs a two-stage semantic matching mechanism to ensure that selected watermark words are genuinely relevant to the document content while maintaining cryptographic security.

Stage 1: Global Semantic Filtering via SECTABLE

For each synthetic document d_i, we first generate its semantic representation using the EMBEDDER:

$$e_{d_i} = \text{EMBEDDER}(d_i) \quad (6)$$

where $e_{d_i} \in \mathbb{R}^d$ captures the document's core semantic features in a high-dimensional embedding space.

The system then computes cosine similarity between e_{d_i} and each vocabulary embedding in SECTABLE to identify candidate watermark words:

$$W_i^{candidate} = \text{Top} * k'(\text{CosineSim}(e * d_i, \text{SECTABLE})) \quad (7)$$

where k' (typically $k' = 2k$) denotes the number of initial candidates selected.The role of this stage is to establish a semantic "shortlist" from the global vocabulary space, narrowing down 3,266 candidate words to a manageable set that shares thematic proximity with the document.

Security Property of SECTABLE The SECTABLE serves as a cryptographic key rather than a simple vocabulary list. It contains 3,266 carefully curated words with pre-computed embeddings that are randomly mapped to the embedding space. This random mapping mechanism ensures that even if attackers identify some watermark words in responses, they cannot reverse-engineer the complete vocabulary structure or predict other watermark words. The mapping remains secure as long as the embedding function and random seed are kept secret.

Stage 2: Semantic Relevance Verification

The candidates from Stage 1 may include semantically irrelevant words due to the random nature of SECTABLE's mapping. To address this limitation, we perform a second-level semantic filtering:

For each candidate word $w \in W_i^{candidate}$, we compute its individual embedding $e_w = \text{EMBEDDER}(w)$ and re-evaluate its semantic relevance to the document:

$$\text{relevance}(w) = \text{CosineSim}(e_w, e_{d_i}) \quad (8)$$

The final watermark vocabulary is then selected as:

$$W_i = \text{Top}_k(\{(w, \text{relevance}(w)) \mid w \in W_i^{candidate}\}) \quad (9)$$

where k represents the target number of watermark words (set to achieve insertion ratio $r = 12\%$).

This two-stage approach achieves a crucial balance: Stage 1 leverages SECTABLE's cryptographic properties for security, while Stage 2 ensures genuine semantic coherence for natural integration and robustness against paraphrasing attacks.

3.3.3. Watermark Embedding Processing for Synthetic Documents

The INSERTER module performs the final watermarking stage by seamlessly integrating the selected watermark vocabulary into synthetic documents through text rewriting. This module operationalizes the post-hoc watermarking paradigm.

Architecture and Operation The INSERTER employs an instruction-following large language model (GPT-4o) that receives the original synthetic document d_s and watermark vocabulary W_i, producing an enhanced version:

$$d_w = \text{INSERTER}(d_s, W_i) \quad (10)$$

The model operates under multi-constraint optimization to ensure: (1) semantic coherence:watermark words fit naturally within context; (2) factual consistency:original claims remain accurate; (3) linguistic fluency:no awkward phrasing; and (4) strategic distribution:words spread across document sections.

Insertion Ratio Control The insertion ratio r balances detection robustness and text quality:

$$r = \frac{|W_i|}{|d_s|} \quad (11)$$

where $|W_i|$ is the number of watermark words and $|d_s|$ is document length. For example, $r = 12\%$ requires inserting 12 watermark words into a 100-word document. Empirical optimization shows that $r = 12\%$ achieves optimal trade-off: lower values compromise detection, while higher values degrade naturalness.

Post-hoc Processing Advantage Unlike generation-time watermarking methods requiring access to model logits, POSTMARK applies watermarks after content creation is complete. This separation ensures: (1) the original dataset D_{do} remains unmodified, (2) watermarking operates independently of generation, and (3) the method applies to any existing text. This design preserves data integrity while achieving robust copyright protection.

3.4. Detection Method

3.4.1. Detection Principle and Technical Architecture

The core principle of the HoneyMark detection mechanism is to leverage the retrieval-generation characteristics of RAG systems to identify watermarked synthetic documents in the corpus. When a RAG system's corpus contains watermarked synthetic documents, specific queries targeting these synthetic contents will trigger the system to retrieve relevant documents and preserve partial watermark vocabulary signals during the response generation process.

The detection task is modeled as a binary hypothesis testing problem[48]:

Null Hypothesis H_0: The RAG system corpus does not contain the target synthetic documents

Alternative Hypothesis H_1: The RAG system corpus contains the target synthetic documents

3.4.2. Targeted Query Generation Strategy

Constructing queries targeted at synthetic documents is one of the crucial components of HoneyMark, utilizing multi-dimensional marker information generated during the synthetic document creation process, including entity markers, temporal markers, geographical markers, and conceptual markers. LLMs are guided through prompts to generate targeted queries based on the marker information and complete synthetic document content.

The generated queries possess dual characteristics: specificity ensures they can effectively trigger the retrieval mechanism of RAG systems containing target synthetic documents; naturalness guarantees that queries will not be identified as probing behavior. Query content is constructed around unique information elements in synthetic documents, which exist only in specific synthetic documents and therefore possess high discriminative power.

3.4.3. Baseline Probability Determination and Statistical Modeling

The baseline probability q_0 is determined through controlled experiments: running the same detection process on RAG systems that are confirmed not to contain watermarked synthetic documents (OUT scenario) and computing the average occurrence rate of the expected watermark vocabulary. This parameter reflects the probability of "coincidental hit" caused by factors such as vocabulary overlap and thematic relevance.

$$q_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{X_i^{\text{OUT}}}{T_i^{\text{OUT}}} \quad (12)$$

where X_i^{OUT} and T_i^{OUT} represent the number of detected vocabulary items and the total number of expected vocabulary items in the i-th query under the OUT scenario, respectively.

Binomial Distribution Model Under the null hypothesis H_0, the occurrence of each vocabulary item in the expected watermark vocabulary list in the response is modeled as independent Bernoulli trials with success probability q_0. Therefore, the total number of detected vocabulary items X follows a binomial distribution:

$$X \mid H_0 \sim \text{Binomial}(T, q_0) \quad (13)$$

3.4.4. Multi-query Aggregation and Statistical Guarantees

Drawing on the aggregation concept of the WARD method, HoneyMark enhances statistical power by accumulating detection results from multiple queries. Given n queries with detection results $(X_1, T_1), (X_2, T_2), \ldots, (X_n, T_n)$, the total detection count is:

$$X_{\text{total}} = \sum_{i=1}^{n} X_i, \quad T_{\text{total}} = \sum_{i=1}^{n} T_i \quad (14)$$

Under the null hypothesis:

$$X_{\text{total}} \mid H_0 \sim \text{Binomial}(T_{\text{total}}, q_0) \quad (15)$$

Using a normal approximation, we construct a standardized test statistic:

$$z = \frac{X_{\text{total}} - T_{\text{total}} \cdot q_0}{\sqrt{T_{\text{total}} \cdot q_0 \cdot (1 - q_0)}} \quad (16)$$

Computing the one-sided p-value:

$$p\text{-value} = 1 - \Phi(z) \quad (17)$$

When z is greater than the significance level α, we reject the null hypothesis and determine that evidence of infringement has been detected.

HoneyMark inherits the statistical rigor of POSTMARK, providing the same level of false-positive control guarantees as WARD. By setting appropriate significance levels, the system can provide legally acceptable statistical evidence proving the probability that the dataset has been used without authorization.

3.5. Robustness Against Data Cleansing Attacks

For any trap-based detection scheme, adversarial data cleansing constitutes a critical threat. A RAG system operator aware of the presence of trap documents may attempt to sanitize their corpus by filtering out documents containing synthetic entities or anomalous content patterns. This section systematically analyzes the robustness of HoneyMark under such attacks.

3.5.1. Intrinsic Robustness of POSTMARK Watermarking

HoneyMark's resilience to cleansing attacks fundamentally stems from the technical properties of the POSTMARK watermarking mechanism. Unlike generation-time watermarks (e.g., KGW red-green watermarks) that embed signals via token-level probability biasing, POSTMARK employs a semantics-driven, post-hoc watermarking approach that exhibits inherent robustness to content manipulation.

Semantics-driven watermark vocabulary selection. POSTMARK selects watermark words by computing the cosine similarity between the semantic embedding of a document and a secret vocabulary table (SECTABLE), ensuring that the chosen words are thematically coherent with the document content. Consequently, these words appear as natural components of domain-specific discourse rather than foreign insertions, making them indistinguishable from legitimate technical terminology under automated linguistic analysis. Empirical evidence shows

that even when explicitly informed of the presence of watermark words, human annotators achieve an F1 score of only 0.06 (precision = 0.46, recall = 0.03) in identifying them [14]. This imperceptibility arises from three design principles: semantic relevance ensures contextual appropriateness, instruction-following LLMs enable natural linguistic integration, and distributed placement avoids clustering artifacts. As a result, standard cleansing heuristics based on "suspicious word frequency" or "lexical anomaly detection" cannot reliably isolate watermark words without incurring prohibitive false positives.

Stability under textual transformation. A decisive advantage of POSTMARK over generation-time watermarks lies in its stability under the multi-stage transformations inherent to RAG workflows. When watermarked documents undergo retrieval, truncation, and generative rewriting processes that typically destroy red-green watermark signals by disrupting token sequences, while POSTMARK's signal persists through semantic preservation. As demonstrated by Chang et al. [14], even under paraphrasing attacks that alter the surface form of watermark words, their semantic association with the document's theme causes synonyms or related concepts to appear in the rewritten text, thereby maintaining statistical detectability. This property directly translates to RAG robustness: when a RAG system retrieves a trap document and synthesizes a response, the semantic relevance of the watermark vocabulary ensures its reappearance in various forms in the output, even if the original phrasing is modified. This explicit, semantics-based signal design (rather than implicit probability biasing) endows POSTMARK with intrinsic stability against both content transformation and rule-based filtering, providing a solid technical foundation for HoneyMark.

3.5.2. Semantic Camouflage of Trap Documents

HoneyMark resists entity- or document-level filtering due to the semantic camouflage achieved by the four-dimensional synthetic document generation framework . Both synthetic entities and complete trap documents are constructed to be indistinguishable from legitimate content under professional scrutiny, rendering effective identification of traps infeasible without incurring unacceptably high false positive rates.

Systematic entity-level camouflage. Our synthetic entities are not randomly generated pseudonyms but carefully constructed conceptual markers that adhere to domain-specific naming conventions and semantic plausibility constraints. The entity synthesis process conditions the generation of core concepts E and associated concepts $\tilde{E}$ on attributes A extracted from real documents in the protected dataset. This ensures that synthetic entities exhibit three key characteristics of real-world entities: (1) **Surface-form conformity**: Entity names follow professional naming patterns (e.g., *"Quantum Dynamics Laboratory"* parallels *"MIT Lincoln Laboratory"*), making them indistinguishable from real institutional names under automated pattern matching; (2) **Domain semantic coherence**: Synthetic entities reside in the same conceptual space as real entities, exhibiting semantic proximity in embedding space; (3) **Contextual logical consistency**: The entity relation network R establishes plausible inter-entity connections, creating context validation chains analogous to real-world knowledge graphs.

The entity linking challenge and the associated high cleansing cost. An adversary might attempt to identify synthetic entities by leveraging entity linking combined with knowledge base

(KB) validation (e.g., checking presence in Wikidata or DBpedia). However, this strategy faces a fundamental coverage–precision trade-off:

1. **Long-tail coverage deficiency**: Even comprehensive KBs suffer from severe long-tail coverage gaps. Studies show that emerging entities, regional organizations, and specialized technical projects in domain-specific corpora routinely lack KB entries despite being legitimate.
2. **"Unlinkable is the norm" phenomenon**: In real-world RAG corpora, particularly those focused on emerging technologies, niche industries, or recent developments, unlinkable entities constitute the majority, not the exception.
3. **Unacceptable deletion trade-off**: Consider a realistic deployment scenario: a corpus of 100,000 documents, of which 5% (5,000 documents) contain legitimate long-tail entities absent from major KBs, and 0.1% (100 documents) are HoneyMark trap documents. Removing all documents containing unlinkable entities would yield: true positives (TP) = 100 trap documents removed; false positives (FP) = 5,000 legitimate documents removed; precision $\approx$ 2%; and a cost ratio of 50:1 (50 legitimate documents deleted per trap removed).

Long-tail documents often represent the competitive advantage of a RAG system, containing proprietary insights, frontier research, emerging trends, and local knowledge. Their removal significantly degrades performance on specialized queries, potentially reducing recall by 20–40% for domain-specific questions targeting long-tail topics. Moreover, deleting interconnected documents disrupts the reference networks, topic clusters, and cross-referential reasoning paths that RAG systems rely on for contextual coherence. Thus, any filtering strategy based on entity "authenticity" forces an untenable trade-off between detection efficacy and system utility.

Document-level indistinguishability. Beyond entity plausibility, complete trap documents achieve professional-grade factual coherence and stylistic naturalness. As shown in Table 1, trap documents maintain a FactScore of 0.95 (indicating 95% of atomic claims are internally consistent) and receive an LLM judge score of 9.0/10. This quality arises from the multi-stage generation pipeline involving attribute-guided synthesis, spatiotemporal grounding, and semantic coherence enforcement. Without prior knowledge of the synthetic entities, automated systems cannot reliably distinguish "potentially fictional content" from "novel but legitimate information." Standard anomaly detection metrics (e.g., topic outliers, stylistic deviation) fail when synthetic documents are explicitly constructed to match corpus-level characteristics.

3.5.3. Statistical Tolerance to Partial Trap Removal

Even if an adversary accepts the operational cost and implements aggressive cleansing, HoneyMark's statistical detection framework exhibits inherent tolerance to partial trap document loss, ensuring reliable infringement detection under adversarial conditions.

HoneyMark's detection mechanism is designed as a multi-query aggregated statistical hypothesis test, not a binary all-or-nothing check. As shown in Table 2, perfect discrimination (AUC-ROC = 1.0) is achieved with only 16 queries, indicating that detection efficacy is more sensitive to watermark signal quality than to the absolute number of trap documents. Under partial trap removal, detection performance degrades gracefully rather than catastrophically: if 50% of trap documents are removed via aggressive cleansing, detection can still succeed by proportionally increasing the number of queries (e.g., 30–40 instead of 20) to accumulate equivalent

statistical evidence. Incomplete cleansing (e.g., removing 20–50% of traps) is insufficient to evade detection, while complete cleansing (removing >90% of traps) requires such aggressive filtering that the RAG system's utility collapses. Consequently, HoneyMark's robustness against data cleansing attacks arises from the synergy of technical design and economic constraints, ensuring practical detectability even under adversarial scrutiny.

4. Experiments and Analysis

4.1. Experimental Setup

Dataset Configuration Our experiments are based on two core datasets to construct realistic RAG infringement detection scenarios. The base corpus employs the Natural Questions (NQ) dataset[49], which contains 307,372 training samples consisting of real user questions submitted to Google Search and their corresponding Wikipedia answers. The protected intellectual property dataset uses News Filtered[50], a large-scale corpus containing content from news websites published under CC BY or CC BY-SA licenses. We assume that a potential semantic overlap exists between the RAG system's base corpus and the protected IP data.

4.2. Implementation Details

Synthetic Document Generation Pipeline : We employ GPT-4o[51] as the core generation model, first conducting in-depth feature analysis of documents in the News Filtered dataset to extract key characteristics including topic distribution, writing style, and content structure. Based on these analytical results, we generate virtual entity relationship networks and spatiotemporal background settings, then integrate this information to compose high-quality synthetic documents. Each generated document maintains stylistic naturalness while ensuring content uniqueness and detectability.

Query Construction and Watermark Embedding : For each synthetic document, we use GPT-4o to construct specialized targeted queries that are carefully designed to be answerable accurately only by RAG systems containing the corresponding synthetic documents. During the watermark embedding phase, GPT-4o is similarly employed to embed POSTMARK watermarks into synthetic documents with a 12% insertion ratio, ensuring semantic robustness and detection reliability of watermark signals.

RAG System Configuration: The experiments employ the LightRAG[52] framework to construct retrieval-augmented generation systems, with the retrieval stage configured to retrieve the top k=3 most relevant documents for each query. This parameter setting ensures retrieval accuracy while controlling computational complexity. In detection experiments, each synthetic document corresponds to one query operation, ensuring efficiency and stealth of the detection process.

Table 1: Factual Consistency Evaluation of Documents Before and After Watermark Embedding

LLM Watermark	FactScoreB	LLM Judge
POSTMARK	0.95	9.0
KGW	0.81	5.6

4.3. Baseline Methods

We select WARD [35] as the primary comparative baseline, representing the current state-of-the-art method in RAG dataset protection. WARD implementation adopts the default configuration from its original paper, using KGW watermarking technology for document rewriting protection. Specific parameter settings include: green list threshold $\gamma = 0.5$, detection threshold $\delta = 3.5$, hash window size $h = 2$. This configuration has been proven in the original WARD paper to achieve a good balance between maintaining text quality and detection performance [35].

We further include AAG (Affirmative Answer-based Guessing) [8] as a representative direct-query membership inference baseline. AAG constructs explicit prompts such as "Does this: <document> appear in your context? Answer with Yes or No." and interprets an affirmative response as evidence of document membership. Anderson et al. [8]demonstrated that AAG achieves high membership inference accuracy on RAG systems without explicit defenses; however, its performance degrades sharply when the system prompt instructs the model to refuse queries about its retrieval context.

To ensure a fair and controlled comparison with HoneyMark, and to isolate the effect of query strategy from retrieval uncertainty, we conduct the AAG evaluation under the same *Def-P* system prompt used in WARD [35] and enforce perfect retrieval: the target document is always included in the top-k retrieved context. This setup eliminates confounding factors from retrieval failure and focuses the evaluation on the inherent detectability and stealth of each method.

4.4. Evaluation Metrics

Factual Consistency: Evaluates the impact of watermark embedding on document factual accuracy. We use FactScore[53] to measure the degree of matching in factual content before and after text watermark embedding. Additionally, we employ GPT-4o as a judge for automated assessment. The evaluation process is as follows: GPT-4o simultaneously receives both original and watermarked versions of documents, scoring their factual consistency on a ten-point scale, with 10 indicating complete consistency.

Detection Performance: We adopt ROC-AUC[54][55] value as the core evaluation metric, which comprehensively measures the detector's discriminative ability to distinguish between infringing and non-infringing RAG systems by analyzing the trade-off relationship between True Positive Rate (TPR) and False Positive Rate (FPR). ROC-AUC values approaching 1 indicate perfect detection performance, while values approaching 0.5 represent random guessing level. Beyond overall AUC metrics, we particularly focus on detection performance under strict false positive rate constraints, reporting key indicators such as TPR@1%FPR. This metric reflects the method's ability to correctly identify infringement behavior when the false alarm rate is strictly controlled below 1% in practical applications, which is of significant importance for legal evidence.

4.5. Experimental Results

4.5.1. Factual Consistency Evaluation

As shown in Table 1, factual consistency quantifies the impact on documents before and after watermark insertion. For synthetic documents generated by our framework, FactScore reaches 0.95, indicating that POSTMARK watermarks cause minimal distortion to factual content. In contrast, KGW-based watermarking (as employed in WARD) yields a substantially lower FactScore of 0.81, reflecting greater semantic deviation introduced by token-level probability manipulation.

To further assess textual fidelity, we employ GPT-4o as an LLM judge to evaluate the factual alignment between original and watermarked versions on a 10-point scale. The POSTMARK-augmented documents receive an average score of 9.0, whereas KGW-rewritten documents score only 5.6. This significant gap underscores that POSTMARK's semantics-driven, post-hoc integration preserves both surface fluency and underlying factual integrity, which is critical for ensuring that trap documents remain indistinguishable from legitimate content in real-world RAG corpora.

4.5.2. Detection Performance

We evaluate the detection performance of HoneyMark on Natural Questions and News Filtered datasets under the experimental configuration described in Section 4.1. As shown in Table 2, HoneyMark achieves perfect discrimination (AUC-ROC = 1.0) with only 16 accumulated queries and attains 100% TPR at 1% FPR with 20 queries.

This dramatic efficiency gain stems from the robustness of POSTMARK's semantics-based watermarking. Unlike WARD's red-green watermarking, which embeds subtle probabilistic biases that are easily diluted during retrieval, truncation, and generative rewriting, POSTMARK embeds semantically relevant vocabulary that persists through the RAG pipeline. Consequently, even a small number of natural, fact-based queries targeting synthetic entities reliably elicit detectable watermark signals in the model's responses.

Moreover, HoneyMark's query strategy, based on unique synthetic facts rather than direct probing, ensures that all interactions remain indistinguishable from legitimate user requests. This not only avoids triggering defensive system prompts (as observed in AAG-style attacks) but also enables consistent signal accumulation across queries. The rapid convergence to perfect classification thus reflects both the technical superiority of POSTMARK and the strategic design of trap documents, establishing HoneyMark as a highly efficient and practical solution for real-world RAG dataset verification.

Table 2: Detection Performance Comparison Under Different Query Numbers

n query	AUC-ROC		TPR@1%FPR	
	ours	WARD	ours	WARD
1	0.8264	0.6216	0.2436	–
2	0.9227	0.6802	0.4556	–
3	0.9633	0.7056	0.6025	–
4	0.9749	0.7338	0.7396	–
5	0.9856	0.7309	0.8174	–
7	0.9960	0.7975	0.9299	–
10	0.9990	0.8505	0.9845	–
12	0.9996	0.8582	0.9950	–
14	0.9997	0.8911	0.9980	–
16	**1.0000**	0.8843	0.9995	–
18	**1.0000**	0.9163	0.9998	–
20	**1.0000**	0.9299	**1.0000**	0.3970
60	**1.0000**	0.9913	**1.0000**	0.7700
80	**1.0000**	**0.9980**	**1.0000**	**0.9180**

4.5.3. Cross-Domain Generalization

To evaluate HoneyMark's robustness in specialized, high-stakes domains, we conduct experiments on the CASESUMM dataset [56], which comprises 25.6K U.S. Supreme Court opinions and their official syllabuses. Legal texts pose a significant challenge for trap-based methods due to their high factual density, strict logical structure, and domain-specific terminology.

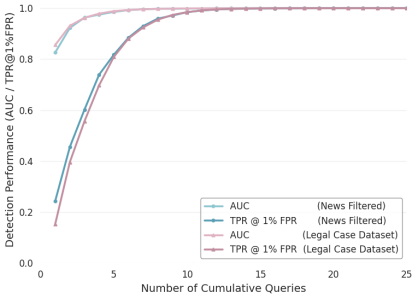


Figure 3: Detection Performance on Legal Case Dataset (CASESUMM)

As shown in Figure 3, HoneyMark achieves perfect discrimination (AUC-ROC = 1.0) with 16–20 queries and reaches 100% TPR at 1% FPR with only 21–25 queries, a performance on par with that on the standard News Filtered dataset (20 queries). This demonstrates that POSTMARK's semantics-driven watermarking remains effective even in highly structured, expert-level corpora. The result is particularly significant for legal practice, where data integrity and copyright verification are critical, and HoneyMark enables non-intrusive, high-confidence detection without compromising the authenticity of original case law.

4.5.4. Robustness Under Resource-Constrained Generation

To assess HoneyMark's practicality in resource-constrained scenarios, we evaluate its performance when trap documents are generated using smaller language models. We conduct experiments using Qwen3-8B (8 billion parameters) for trap document synthesis and watermark embedding, while maintaining GPT-4o as the RAG system's response model. This asymmetric configuration simulates a realistic scenario where copyright holders operate under budgetary constraints but must verify compliance of well-resourced RAG deployments.

As shown in Figure 4, Qwen3-8B-generated traps require 40–50 queries to achieve perfect discrimination (AUC-ROC = 1.0) and 100% TPR at 1% FPR, compared to 16–20 queries for GPT-4o-generated traps. This degradation stems from two factors: reduced linguistic sophistication in synthetic entity construction may marginally decrease retrieval frequency, and weaker instruction-following during watermark insertion occasionally produces less natural vocabulary integrations, attenuating signal strength per query. Critically, both configurations ultimately converge to perfect classification, indicating that the performance gap manifests as query efficiency rather than fundamental detectability failure. Despite requiring a 2–2.5 fold increase in the number of queries, the total query count remains within a practical range for effective verification.

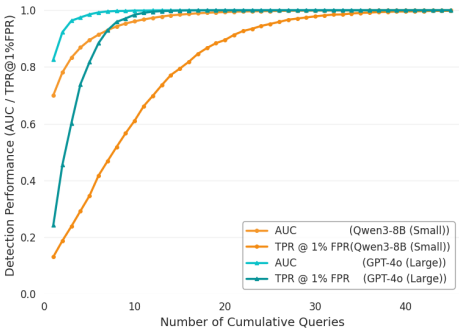


Figure 4: Detection Performance: Small vs. Large Model Generation

4.5.5. Stealth Evaluation Against Direct-Query Baselines

To assess HoneyMark's operational stealthiness in adversarial environments, we conduct a controlled comparison against the AAG baseline under defensive conditions. Operating against

the same RAG system configured with the Def-P defensive prompt that instructs the model to refuse context-related inquiries. Perfect retrieval is enforced to isolate query detectability from retrieval uncertainty.

As shown in Figure 5, the two methods exhibit fundamentally different behavior under defensive conditions. AAG suffers catastrophic failure with 87.0% of queries triggering defensive refusals such as "unanswerable" or explicit rejection responses, while only 9.0% successfully elicit valid membership signals, rendering it practically unusable in protected environments. In stark contrast, HoneyMark achieves perfect stealth with 0% defense triggering and 94.2% of queries producing detectable watermark signals. The remaining 5.8% represent natural vocabulary attenuation during generation rather than defensive blocking. This dramatic difference stems from HoneyMark's semantic-level indistinguishability - its entity-focused queries like "Explain the development of Monarchical Symbolism in democratic societies" are structurally and semantically identical to legitimate user requests, effectively bypassing instruction-based guardrails. This enables copyright verification without exposing detection intent, a crucial capability for compliance audits where covertness determines operational feasibility.

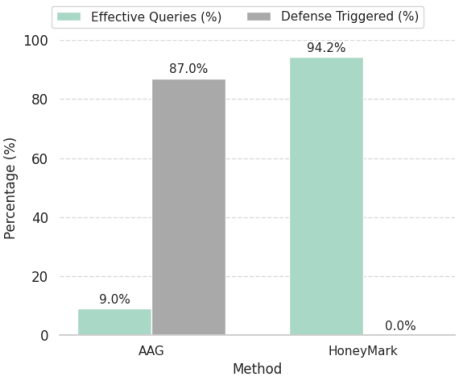


Figure 5: Comparison of Effective Query Rate and Defense Triggering Rate for AAG vs. HoneyMark under Def-P

5. Conclusion

This paper addresses the critical challenge of dataset copyright verification in RAG systems through HoneyMark, a trap document-based detection method that achieves reliable verification without modifying original data. By combining a four-dimensional synthetic document generation framework with POSTMARK post-hoc watermarking, HoneyMark preserves complete data integrity while enabling statistical detection of unauthorized dataset usage.

HoneyMark offers three key advantages over existing approaches: (1) Zero modification: Original datasets remain completely unchanged, eliminating factual distortion risks inherent in rewriting-based methods; (2) Pipeline robustness: POSTMARK's semantics-driven watermarks demonstrate superior stability through RAG's multi-stage processing (retrieval, truncation, and generation) compared to generation-time watermarks that rely on fragile token-level probability biases; (3) Detection efficiency: Synthetic entities with unique characteristics enable targeted queries that achieve perfect discrimination.

Comprehensive experiments validate HoneyMark's effectiveness across diverse scenarios. On standard news datasets, the method achieves perfect discrimination (AUC-ROC = 1.0) with only 16 queries and 100% TPR at 1% FPR with 20 queries, significantly outperforming WARD which requires 80+ queries for comparable performance. Cross-domain validation on legal case corpora demonstrates consistent efficacy, while stealth evaluation shows zero defensive triggering compared to 87% for direct-query baselines. Even under resource-constrained generation

(Qwen3-8B), the method maintains perfect classification with 40–50 queries, confirming practical viability across deployment scenarios.

6. Limitations

While HoneyMark demonstrates significant advantages in RAG dataset verification, several limitations merit discussion.

Dependence on High-Quality Language Models The framework's performance depends on the generation capabilities of instruction-following LLMs. As shown in Section 4.5.4, smaller models (e.g., Qwen3-8B) require 2–2.5× more queries than GPT-4o to achieve equivalent detection performance, creating a trade-off between computational cost and detection efficiency.

Adversarial Data Cleansing Although Section 3.5 demonstrates that semantic camouflage and POSTMARK's stability create prohibitive filtering costs, a determined adversary combining entity linking with manual inspection could potentially remove subsets of trap documents. The statistical framework degrades gracefully under partial removal but comprehensive evaluation against coordinated multi-stage attacks remains future work.

Domain Generalization While experiments on legal corpora (Section 4.5.3) validate cross-domain effectiveness, specialized fields with strict terminological conventions (e.g., medical records, financial reports) may require domain-specific tuning of the generation framework.

Dynamic Dataset Scenarios The current design targets static or periodically updated datasets. Continuously evolving corpora require additional mechanisms for incremental trap generation and adaptive watermark refreshing to maintain detection efficacy.

Despite these limitations, HoneyMark's core contributions, including zero-modification verification, semantics-driven watermarking, and statistical guarantees, provide a robust foundation for practical RAG dataset protection.

Acknowledgments

This work is supported by the National Key R&D Program of China (2024YFF0907000), which is also the research achievement of the Key Laboratory of Digital Right Services of National Press and Publication Administration.

References

[1] P. Lewis, *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2021, arXiv:2005.11401.

[2] S. Gupta, R. Ranjan and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," 2024, arXiv:2410.12837.

[3] Y. Gao, *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," 2024, arXiv:2312.10997.

[4] J. Yu, *et al.*, "MIND-RAG: Multimodal Context-Aware and Intent-Aware Retrieval-Augmented Generation for Educational Publications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Honolulu, HI, USA, Oct. 19–20, 2025, pp. 4216–4223, doi: 10.1109/ICCVW69036.2025.00438.

[5] M. A. Brown, A. Gruen, G. Maldoff, S. Messing, Z. Sanderson and M. Zimmer, "Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations," 2024, arXiv:2410.23432.

[6] A. Golatkar, A. Achille, L. Zancato, Y.-X. Wang, A. Swaminathan and S. Soatto, "CPR: Retrieval Augmented Generation for Copyright Protection," 2024, arXiv:2403.18920.

[7] P. Adler, *et al.*, "Auditing Black-box Models for Indirect Influence," 2016, arXiv:1602.07043.

[8] M. Anderson, G. Amit and A. Goldsteen, "Is My Data in Your Retrieval Database? Membership Inference Attacks Against Retrieval Augmented Generation," in *Proc. 11th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP)*, Porto, Portugal, Feb. 20–22, 2025, pp. 474–485, doi: 10.5220/0013108300003899.

[9] N. Jovanović, R. Staab, M. Baader and M. Vechev, "Ward: Provable RAG Dataset Inference via LLM Watermarks," presented at the *13th Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 24–28, 2025.
[10] Z. Cai, S. Liu, H. Wang, H. Zhong and X. Li, "Towards Better Statistical Understanding of Watermarking LLMs," 2024, arXiv:2403.13027.
[11] Y. Liu, X. Zhao, D. Song and Y. Bu, "Dataset Protection via Watermarked Canaries in Retrieval-Augmented LLMs," 2025, arXiv:2502.10673.
[12] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers and T. Goldstein, "A Watermark for Large Language Models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, Jul. 23–29, 2023, pp. 17061–17084.
[13] J. Su, J. P. Zhou, Z. Zhang, P. Nakov and C. Cardie, "Towards More Robust Retrieval-Augmented Generation: Evaluating RAG Under Adversarial Poisoning Attacks," 2025, arXiv:2412.16708.
[14] Y. Chang, K. Krishna, A. Houmansadr, J. F. Wieting and M. Iyyer, "PostMark: A Robust Blackbox Watermark for Large Language Models," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Miami, FL, USA, Nov. 12–16, 2024, pp. 8969–8987, doi: 10.18653/v1/2024.emnlp-main.506.
[15] E. German, S. Antebi, E. Habler, A. Shabtai and Y. Elovici, "LexiMark: Robust Watermarking via Lexical Substitutions to Enhance Membership Verification of an LLM's Textual Training Data," 2025, arXiv:2506.14474.
[16] J. Ren, *et al.*, "A Robust Semantics-Based Watermark for Large Language Models Against Paraphrasing," in *Findings Assoc. Comput. Linguistics: NAACL 2024*, Mexico City, Mexico, Jun. 16–21, 2024, pp. 613–625, doi: 10.18653/v1/2024.findings-naacl.40.
[17] Y. Li, G. Liu, C. Wang and Y. Yang, "Generating Is Believing: Membership Inference Attacks Against Retrieval-Augmented Generation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Hyderabad, India, Apr. 6–11, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889013.
[18] M. Liu, S. Zhang and C. Long, "Mask-based Membership Inference Attacks for Retrieval-Augmented Generation," 2025, arXiv:2410.20142.
[19] B. An, *et al.*, "WAVES: Benchmarking the Robustness of Image Watermarks," 2024, arXiv:2401.08573.
[20] A. Liu, *et al.*, "A Survey of Text Watermarking in the Era of Large Language Models," 2024, arXiv:2312.07913.
[21] Y. Liang, J. Xiao, W. Gan and P. S. Yu, "Watermarking Techniques for Large Language Models: A Survey," 2024, arXiv:2409.00089.
[22] P. Lv, *et al.*, "RAG-WM: An Efficient Black-Box Watermarking Approach for Retrieval-Augmented Generation of Large Language Models," 2025, arXiv:2501.05249.
[23] J. Kirchenbauer, *et al.*, "On the Reliability of Watermarks for Large Language Models," 2024, arXiv:2306.04634.
[24] X. Wang, *et al.*, "Building Intelligence Identification System via Large Language Model Watermarking: A Survey and Beyond," 2024, arXiv:2407.11100.
[25] S. Dathathri, *et al.*, "Scalable Watermarking for Identifying Large Language Model Outputs," *Nature*, vol. 634, no. 8035, pp. 818–823, Oct. 2024, doi: 10.1038/s41586-024-08025-4.
[26] J. Zhang, *et al.*, "CoheMark: A Novel Sentence-Level Watermark for Enhanced Text Quality," 2025, arXiv:2504.17309.
[27] S. B. Jabra and M. A. B. Farah, "Deep Learning-Based Watermarking Techniques Challenges: A Review of Current and Future Trends," *Circuits Syst. Signal Process.*, vol. 43, no. 7, pp. 4339–4368, Jul. 2024, doi: 10.1007/s00034-024-02651-z.
[28] Y. Huang, H. Guan, J. Liu, S. Zhang, B. Niu and G. Zhang, "Robust Texture-Aware Local Adaptive Image Watermarking With Perceptual Guarantee," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4660–4674, Sep. 2023, doi: 10.1109/TCSVT.2023.3245650.
[29] L. Spitzner, *Honeypots: Tracking Hackers*. Boston, MA, USA: Addison-Wesley, 2002.
[30] B. M. Bowen, S. Hershkop, A. D. Keromytis and S. Stolfo, "Baiting Inside Attackers Using Decoy Documents," in *Proc. 5th Int. ICST Conf. Secur. Privacy Commun. Netw. (SecureComm)*, Athens, Greece, 2009, pp. 51–70, doi: 10.1007/978-3-642-05284-2_4.
[31] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, May 22–24, 2017, pp. 3–18, doi: 10.1109/SP.2017.41.
[32] J. Franco-Contreras and G. Coatrieux, "Robust Watermarking of Relational Databases With Ontology-Guided Distortion Control," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 9, pp. 1939–1952, Sep. 2015, doi: 10.1109/TIFS.2015.2439962.
[33] M. Shehab, E. Bertino and A. Ghafoor, "Watermarking Relational Databases Using Optimization-Based Techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 116–129, Jan. 2008, doi: 10.1109/TKDE.2007.190668.
[34] M. L. P. Gort, M. Olliaro, A. Cortesi and C. F. Uribe, "Semantic-Driven Watermarking of Relational Textual Databases," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114013, doi: 10.1016/j.eswa.2020.114013.
[35] S. Ge, Y. Guo, H. Zhang, L. Zhang and X. Zhang, "PSOHSW: A Robust Database Watermarking Scheme With Low Distortion," in *Proc. 4th Int. Conf. Cryptography, Netw. Secur. Commun. Technol. (CNSCT)*, Zhengzhou, China, Jan. 17–19, 2025, pp. 50–54, doi: 10.1145/3723890.3723900.

[36] H. He, J. Zhang and F. Chen, "Fuzzy Database Mining and Information Retrieval: A Chaos-Based Fragile Watermarking Scheme With Recovery," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Haikou, Hainan, China, Aug. 24–27, 2007, vol. 2, pp. 303–307, doi: 10.1109/FSKD.2007.324.

[37] G. Majumder, P. Pakray, A. Gelbukh and D. Pinto, "Semantic Textual Similarity Methods, Tools, and Applications: A Survey," *Comput. Sist.*, vol. 20, no. 4, pp. 647–665, Dec. 2016, doi: 10.13053/CyS-20-4-2506.

[38] F. Benedetti, D. Beneventano, S. Bergamaschi and G. Simonini, "Computing Inter-Document Similarity With Context Semantic Analysis," *Inf. Syst.*, vol. 80, pp. 136–147, Feb. 2019, doi: 10.1016/j.is.2018.02.009.

[39] B. Ni, *et al.*, "Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey," 2025, arXiv:2502.06872.

[40] A. Liu, L. Pan, X. Hu, S. Meng and L. Wen, "A Semantic Invariant Robust Watermark for Large Language Models," 2024, arXiv:2310.06356.

[41] M. Huo, S. A. Somayajula, Y. Liang, R. Zhang, F. Koushanfar and P. Xie, "Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models," 2024, arXiv:2402.18059.

[42] Y. Liu, T. Cong, M. Backes, Z. Li and Y. Zhang, "Watermarking LLM-Generated Datasets in Downstream Tasks," 2025, arXiv:2506.13494.

[43] K. Ning, *et al.*, "MCGMark: An Encodable and Robust Online Watermark for Tracing LLM-Generated Malicious Code," 2025, arXiv:2408.01354.

[44] S. Abdelnabi and M. Fritz, "Adversarial Watermarking Transformer: Towards Tracing Text Provenance With Data Hiding," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Virtual Conf., May 24–27, 2021, pp. 121–140, doi: 10.1109/SP40001.2021.00083.

[45] S. Herbold, "Semantic similarity prediction is better than other semantic similarity measures," 2024, arXiv:2309.12697.

[46] Z. Fei, B. Yi, J. Geng, R. He, L. Nie and Z. Liu, "Your Semantic-Independent Watermark is Fragile: A Semantic Perturbation Attack against EaaS Watermark," 2025, arXiv:2411.09359.

[47] X. Li, F. Ruan, H. Wang, Q. Long and W. J. Su, "A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules," *Ann. Statist.*, vol. 53, no. 1, pp. 322–351, Feb. 2025, doi: 10.1214/24-AOS2468.

[48] B. Huang, *et al.*, "Towards Optimal Statistical Watermarking," 2024, arXiv:2312.07930.

[49] T. Kwiatkowski, *et al.*, "Natural Questions: A Benchmark for Question Answering Research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 453–466, Nov. 2019, doi: 10.1162/tacl_a_00276.

[50] N. Kandpal, *et al.*, "The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text," 2025, arXiv:2506.05209.

[51] A. Hurst, *et al.*, "GPT-4o System Card," 2024, arXiv:2410.21276.

[52] Z. Guo, L. Xia, Y. Yu, T. Ao and C. Huang, "LightRAG: Simple and Fast Retrieval-Augmented Generation," 2024, arXiv:2410.05779.

[53] S. Min, *et al.*, "FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation," 2023, arXiv:2305.14251.

[54] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[55] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.

[56] M. Heddaya, K. MacMillan, A. Malani, H. Mei and C. Tan, "CaseSumm: A Large-Scale Dataset for Long-Context Summarization from U.S. Supreme Court Opinions," 2024, arXiv:2501.00097.