2024 Annual Computer Security Applications Conference Workshops (ACSAC Workshops) | 979-8-3315-3281-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACSACW65225.2024.00010

# Dark Watchdog: A Novel RAG-Driven System for Real-Time Detection and Analysis of Data Leaks on Dark Web Forums

Shing-Li Hung
*Institue of Information Security*
*National Tsing Hua University*
*Hsinchu, Taiwan*

Chung-Kuan Chen
*CyCraft Technology*
*CyCraft Technology*
*New Taipei City, Taiwan*

Keisuke Furumoto
*Cybersecurity Research Institute*
*National Institute of Information*
*and Communications Technology*
*Tokyo, Japan*

Takeshi Takahashi
*Cybersecurity Research Institute*
*National Institute of Information*
*and Communications Technology*
*Tokyo, Japan*

Hung-Min Sun
*Departmant of Computer Science*
*National Tsing Hua University*
*Hsinchu, Taiwan*

***Abstract*—Personal data breaches have become increasingly common, making the dark web a key marketplace for trading stolen information, often without the immediate awareness of affected organizations. To address this challenge, we introduce Dark Watchdog, a novel system that actively monitors dark web forums and employs a specially fine-tuned BERT classification model to categorize transaction posts into five distinct types of breaches with high accuracy. Dark Watchdog uniquely integrates retrieval-augmented generation (RAG) to efficiently vectorize and analyze dark web data, allowing cyber security analysts to access the latest intelligence on data leaks while preserving privacy by minimizing data exposure to large language models (LLMs). This approach not only improves detection precision but also optimizes computational resources by reducing token usage. Dark Watchdog offers an innovative and practical solution for real-time dark web monitoring, enabling timely insights into ongoing data leak incidents and enhancing the overall effectiveness of cyber security efforts.**

***Index Terms*—Dark web, cyber security, Tor, BERT, Classification, Retrieval-augmented generation, RAG, Large language model, LLM, Privacy**

## 1. Introduction

In recent years, there has been an increasing trend of ransomware groups stealing organizations' private information to use as leverage for extortion or employing phishing schemes. [1]. A primary reason for the occurrence of fraud cases is the leakage of personal data [2], which allows criminals to gain detailed knowledge of the victims and customize their fraudulent schemes. To prevent such incidents, it is crucial to control the leakage of data at the source, which relies on the active cooperation of companies and government agencies handling personal data. However, incidents of data leakage continue to occur.

Leaked personal data are disseminated through buying and selling, where attackers obtain personal information through vulnerabilities or internal leaks and then sell it on the internet to other interested buyers. For both buyers and sellers, "anonymity" is a critical and necessary technology, and the dark web provides anonymity, which makes it the most important medium for the trade of personal data.

The dark web cannot be accessed via standard search engines and is only accessible through networks such as Tor or I2P, with URLs ending in .onion [3], [4]. The dark web hosts a variety of websites, including ransomware, forums, and social media platforms [5]–[8]. Our research identified the largest forum for the sale of personal data: BreachForums. This site has a large user base that discusses and trades various types of personal data, with frequent updates to posts daily. Therefore, this paper focuses on monitoring this site for posts related to the sale of personal data connected to Taiwan.

Given the large number of daily posts and the variety of leaked data types, such as personal-data sales, account passwords, and leaked game-source codes, we collected posts from BreachForums for deep-learning training. We aimd to train a classification model to accelerate analysts' understanding of the posts.

Since the collected data may contain personal information, we used retrieval-augmented generation (RAG) to vectorize the organized data. These data were then encapsulated into prompts of interest using the latest large language model (LLM) to assist researchers in rapidly analyzing data on the dark web.

The following summarizes the contributions of this paper:

979-8-3315-3281-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ACSACW65225.2024.00010

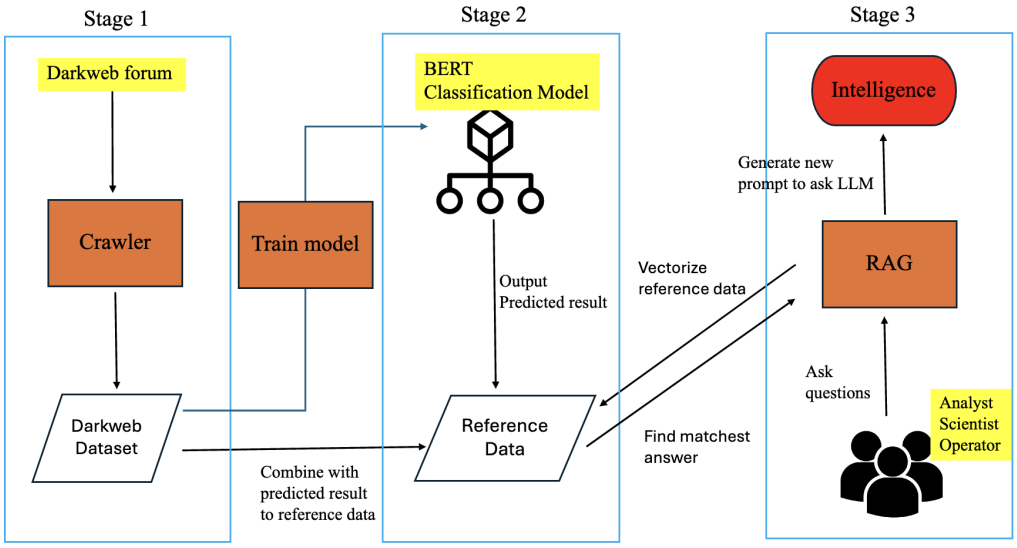


Figure 1. The Architecture of Dark Watchdog

- We independently collected data from dark web forums and organized it into a dataset containing post articles and content. This dataset was used to train a BERT model specialized for classification.
- Our BERT model is capable of detecting five different types of data-leak events. This is an advancement over previous studies that applied to post classification [9]; our model performs better in metrics such as accuracy and precision. Our model provides more focused detection of leak events, offering five distinct types, and exhibits good performance in terms of accuracy and confusion matrix.
- We used an LLM to analyze the organized data from dark web posts, comparing the differences in the model's responses with and without the use of RAG. Ultimately, without exposing our dark web data, we effectively reduced the number of tokens in our prompts.

In recent years, numerous academic studies have begun to focus on data from the dark web, leading to a proliferation of research on data collection methods. Lawrence et al. introduced D-miner in 2017 [10], a modular framework designed to allow users to crawl data from the dark web and parse it into JSON format. Schäfer et al. proposed BlackWidow in 2019 [11], a Docker-based microservice for monitoring the dark web. ZHENG et al. [12] in 2023 shared a system design for an efficient dark web crawler that efficiently acquires data from the dark web.

Zenebe et al. attempted to classify posts from the dark web, similar to our intended task; however, they employed traditional machine learning methods, which did not achieve high accuracy [9]. In our paper, we aim to improve upon their results by using more advanced techniques to enhance classification performance.

Vaswani et al. introduced the Transformer architecture with their seminal work "Attention is all you need" [13], which has since been widely adopted in the field of Natural Language Processing. A year later, the Bidirectional Encoder Representations from Transformers, known as BERT [14], was released. In 2023, Jin et al. proposed DarkBERT [15], marking the first application of BERT to dark web data. However, for the sake of generalizability, it was not trained to specialize, hence this paper aims to fine-tune a BERT model specialized for classifying data leak tasks.

Sultana et al. have compiled recent research on the use of Large Language Models in cybersecurity automation [16].

## 2. Background

### 2.1. Dark web

The online realm is segmented into the surface web and deep web. The surface web comprises web pages that are accessible through standard search engines and serves as the primary mode for information retrieval by the general populace. In contrast, the deep web encompasses content not indexable by standard search engines, including personal shopping carts, online banking, and subscription-based streaming services. The dark web, a subset of the deep web, requires special software configurations for access. Accessing the dark web typically involves the use of The Onion Router (Tor) network [3], [4], which is an anonymity network operated by volunteers. This network enables users to browse the web while concealing their Internet Protocol addresses and identities, thereby safeguarding their privacy and security. By connecting through the Tor network, users can access websites with the .onion suffix, which are domains within the dark web.

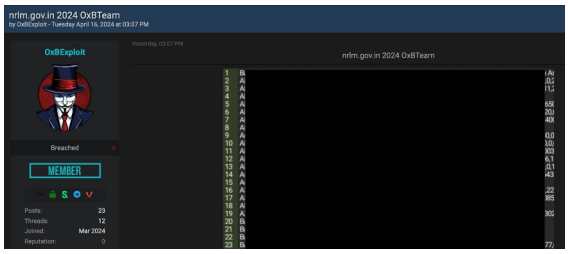


Figure 2. BreachForums website

## 2.2. BreachForums

BreachForums (is shown in Figure 2) is an online forum on the dark web that functions as a marketplace. For cybercriminals, BreachForums provides a platform to trade personal privacy data, databases, credit card information, and hacking tools. Previously, the largest data leak site in the world was RaidForums. However, following its shutdown by the FBI, BreachForums has emerged as the new largest dark web marketplace for such transactions [17].

## 2.3. BERT

Vaswani et al. introduced the Transformer architecture in their seminal work [13], which has since been widely adopted in the field of NLP. A year later, BERT [14], was released.

BERT represents a significant advance in the field of natural language processing (NLP). BERT has been pretrained on a large corpus of text using two primary tasks: masked language modeling (MLM) and next sentence prediction (NSP). In MLM stage, random words in a sentence are replaced with a placeholder, and the model is trained to predict the original words on the basis of the context provided by the other words in the sentence. The NSP task involves determining whether a given sentence logically follows another sentence, which helps the model determine the relationships between consecutive sentences.

We have the base BERT model, now we need a custom dataset that we can further train the model on specific tasks for specialized capabilities.

Figure 3 shows how to perform the classification task using the BERT architecture

- Input Stage: The text is first broken down into multiple tokens (Token 1 to Token n), each representing a word or character within the text. A special token [CLS] is added at the beginning of the input sequence. This token is used in BERT training for classification tasks, and its final hidden state represents the self-attention information of the entire input sequence.
- BERT Processing Stage: All tokens, including [CLS], are fed into BERT. BERT processes these tokens through multiple layers of the transformer architecture, where each layer further encodes the contextual information of the text on the basis of the previous layer. Each layer of BERT uses self-attention to analyze the relationships between tokens, thereby capturing the contextual features of the entire text.
- Output Stage: The output of BERT consists of the hidden states corresponding to each token, with the hidden state of the [CLS] token commonly used to represent the entire sequence. This hidden state is then passed to a Softmax layer or another classifier, which transforms the representation of the [CLS] token into a final class prediction. The function of the Softmax layer is to calculate the probabilities for each class and select the class with the highest probability as the output result.

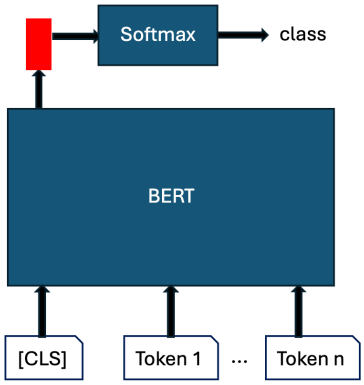


Figure 3. BERT classification task

## 2.4. Retrieval-augmented generation

RAG is an NLP technique that integrates retrieval and generative technologies with the aim of enhancing the quality and relevance of responses from LLM. Figure 4 illustrates the architecture of RAG. When a human user asks a question, RAG begins by vectorizing the reference data. It then identifies the documents or segments within the reference vectors that are most closely related to the query, a process known as the retrieval phase. The identified vectors are subsequently recombined with the original question to form a new prompt. In the final stage, known as generation, this prompt is submitted to the used LLM, which generates an answer. This approach allows the model to use specific reference background data when generating responses, improving the accuracy and richness of the answers.

Benefits of RAG:

- Dynamic knowledge base: Leverages up-to-date information from external sources, which is particularly valuable in rapidly evolving fields such as cyber security or computer science.
- Reduce the number of tokens: Compared with attaching all the reference data, it can significantly reduce the number of tokens. It can also enhance

TABLE 1. METADATA DESCRIPTION OF DARK WEB DATASET

| Field | Description |
|---|---|
| State | Defined by website |
| Title | Post's title |
| Ctime | Create time |
| Replies | Number |
| Views | Number |
| Content | Post itself (not include reply) |
| Page type | Board's name (category) |

TABLE 2. METADATA DESCRIPTION OF REFERENCE DATA

| Field | Description |
|---|---|
| Subboard | Board of post |
| Title of article | Title of Post |
| Is related to Taiwan | Use RE to target "taiwan" or "tw" keyword |
| 5 labels | Output of our BERT classification model |

processing speed and reduce operational costs, particularly when using token-based pricing models of LLMs.

- Maintain privacy: One does not need to submit all reference data to an LLM company. This approach significantly mitigates the risk of data breaches and unauthorized accesses, making RAG an ideal choice for industries in which data security is important.

In 2023, Jin et al. proposed DarkBERT [15], marking the first application of BERT to dark web data. However, for the sake of generalizability, it was not trained to specialize; hence, we fine-tuned a BERT model specialized for classifying data-leak tasks.

## 3. Methodology

The methodology of this research is divided into three main parts, as illustrated in Figure 1

1. Collection of data from targeted dark web forums and organizations into a labeled dataset.

2. Training a BERT classification model to categorize full-text data into five distinct types of leaks.

3. Using RAG to allow our reference data to be queried by an LLM, acquiring daily dark web intelligence, and maintaining privacy.

We initially focused on the BreachForums website for data scraping, obtaining data related to "posts" and first organizing the necessary metadata, as referred to in Table 1. A custom field we created, "Page type", is derived from the category of the board to which a post belongs. Given that we selected boards related to "leaks", and BreachForums is the world's largest personal data trading forum, the leak types are already categorized, enabling us to collect board names corresponding to the posts.

Upon investigation, BreachForums was found to host 33 different boards, encompassing personal-data sales, global news, and technology discussions, among others, presenting a diverse forum. We focused on detecting posts selling personal data; thus, we selected four boards related to leakage and categorized unrelated posts, such as global news and technology discussions, as "Other".

The boards are defined as follows:

- Cracked-Accounts: Accounts where security has been breached, with credentials cracked or decrypted, potentially from various online services such as social media, streaming, or banking platforms.
- Combolists: These are collections of compiled usernames and passwords from various sources. Combolists are often used by cybercriminals to carry out credential stuffing attacks, where automated tools are used to access multiple user accounts across different platforms.
- Databases: Large collections of structured data that have been compromised, including user account details, personal information, transaction histories, etc., typically from a single source or platform.
- Stealer-Logs: Records from malware designed to steal information from infected computers, including keystrokes, browser history, screenshots, and other personal data.
- Other: Posts unrelated to leaks on BreachForums, including global news, technology discussions, music discussions, etc., since this model can be applied to any type of post.

Ultimately, our dataset comprised five different board categories, totaling 10,901 dark web posts. We then vectorized this data using term frequency–inverse document frequency (TF-IDF), reduced dimensions with the t-distributed stochastic neighbor embedding (t-SNE) algorithm, and projected it onto a two-dimensional plane, as shown in Figure 6. This clear segmentation into five distinct categories validated the rationality of our selected categories.

Next, we prepared the labeled dataset by selecting suitable features, combining the "title" and "content" columns into a new column called "text", with each post labeled according to its board. This prepared dataset facilitated the training of a BERT classification model.

Classification: Since the post content is entirely in English, this constitutes an NLP classification task. We used BERT as the foundational model for classification, training it with data from the dark web forum. Given BreachForums' focus on leaked personal data, this model could also be applied to other forum types, such as Dread [18] or X(formerly Twitter) [19] because these websites all have post types data.

We split the data into training and testing sets in a 7:3 ratio. During the training stage, we used the Early Stopping method, where the training is halted when the accuracy on the validation dataset starts to decrease, in order to avoid overfitting.

RAG: We organized information from the previous dataset and new fields as reference material for our RAG tasks. We selected the board and title of posts, used regular expressions to identify any keywords related to Taiwan, and included classification results from the first model phase. This setup is detailed in Table 2 and Figure 5.

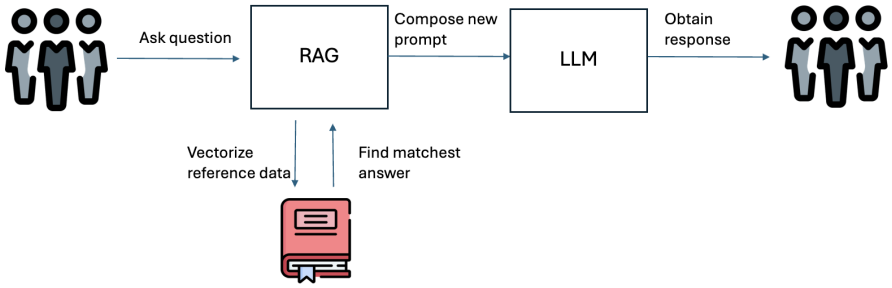


Figure 4. RAG architecture

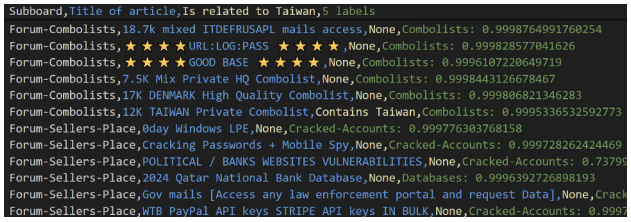


Figure 5. Reference data

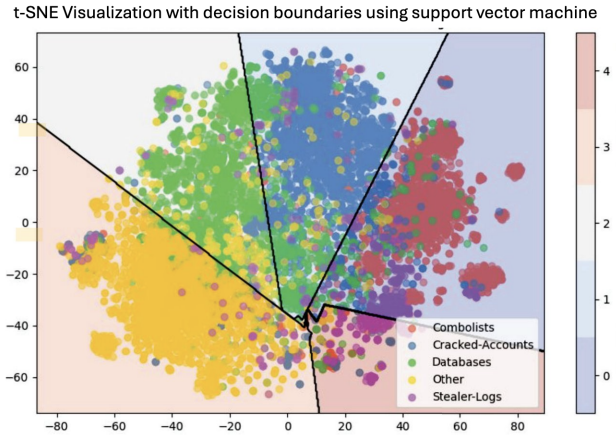


Figure 6. t-SNE dimensionality reduction

Although the target is specifically focused on Taiwan-related strings, Dark Watchdog allows flexible adjustments to the regular expressions, making it applicable to other countries or regions as well.

LangChain is an open-source framework designed to assist developers in integrating large language models (LLMs) with various external tools and data sources. Its key features include an LLM interface, prompt templates, agents, and retrieval modules.

Using the LangChain framework, we performed the RAG tasks, extracting daily updates from dark web intelligence. This enabled us to query the LLM in natural language about the security risks present in the referenced data and other issues relevant to cybersecurity analysts.

# 4. Evaluation

## 4.1. BERT classification model

The evaluation of the model in terms of its classification performance is detailed as follows. We set the number of epochs to 10, successfully reducing the final training loss to 0.0028. Table 3 shows that the model achieved an accuracy of 94.177%, with an area under the receiver operating characteristic curve (ROC AUC) score of 96.242%. Overall, the classification results are satisfactory. The effective segmentation of the data into five predetermined categories during the initial dimension reduction using the t-SNE algorithm anticipated these favorable outcomes.

The performance metrics for each category are listed in Table 4

Noted that the ”Stealer Logs” category, which had fewer posts, showed poorer performance. However, even the lowest-performing category achieved a recall of 89%, indicating that the classification task performed robustly across different types of data. Figure 8 presents the model’s confusion matrix. Except for the ”Stealer Logs,” which are lighter in color due to fewer numbers, all categories are well distinguished, signifying accurate predictions across the dataset, with the model successfully identifying the correct categories.

Next, we compared the previous academic research with our BERT model in Table 3. Zenebe et al. [9] used a machine learning model to classify dark web posts. They used three models: Random Forest, Decision Tree, and Naive Bayes, and provided a detailed evaluation of each model’s accuracy, ROC AUC score, precision, recall and F1 score. Our BERT model demonstrated superior performance in accuracy, precision, recall and F1 score when compared with these models.

Since this dataset comes from a real-world environment, it is influenced by many outliers. During the calculation of the ROC curve, extreme data points or outliers can impact the results. If the BERT model performs particularly poorly on these outliers, it could lead to results that are slightly lower compared to other ML models.

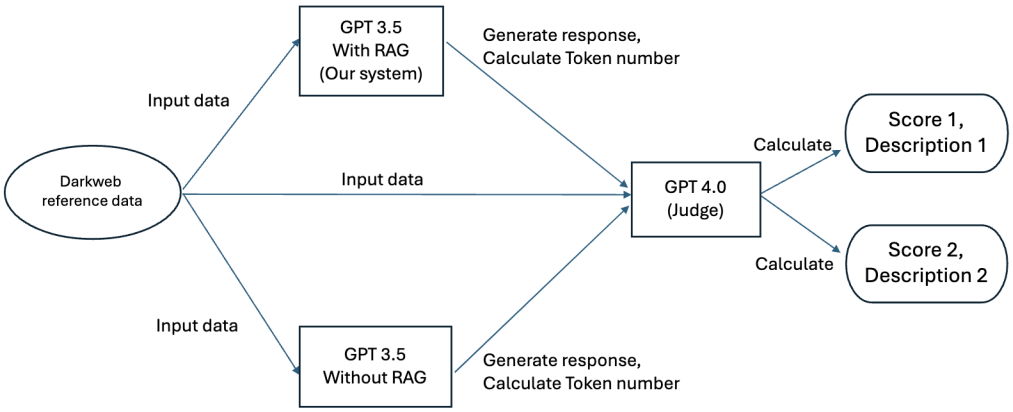


Figure 7. RAG evaluation architecture

TABLE 3. MODEL COMPARISON

| Model | Accuracy (%) | ROC (%) | Precision (%) | Recall (%) | F1 score (%) | Reference |
|---|---|---|---|---|---|---|
| BERT | 94.18 | 96.24 | 94.18 | 94.18 | 94.17 | This study |
| Random Forest | 88.26 | 97.83 | 87.41 | 74.20 | 75.82 | [9] |
| Decision Tree | 84.87 | 87.62 | 80.17 | 78.24 | 79.12 | [9] |
| Naive Bayes | 86.61 | 96.27 | 90.68 | 70.26 | 72.17 | [9] |

TABLE 4. CLASSIFICATION RESULTS OF EACH CATEGORY

| **Class** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| Cracked-Accounts | 0.95 | 0.95 | 0.95 | 603 |
| Databases | 0.92 | 0.94 | 0.93 | 450 |
| Combolists | 0.96 | 0.97 | 0.96 | 464 |
| Stealer-Logs | 0.93 | 0.89 | 0.91 | 141 |
| Other | 0.94 | 0.93 | 0.93 | 523 |
| **Accuracy** | | 0.94 (2181) | | |
| **Macro avg** | 0.94 | 0.94 | 0.94 | 2181 |
| **Weighted avg** | 0.94 | 0.94 | 0.94 | 2181 |

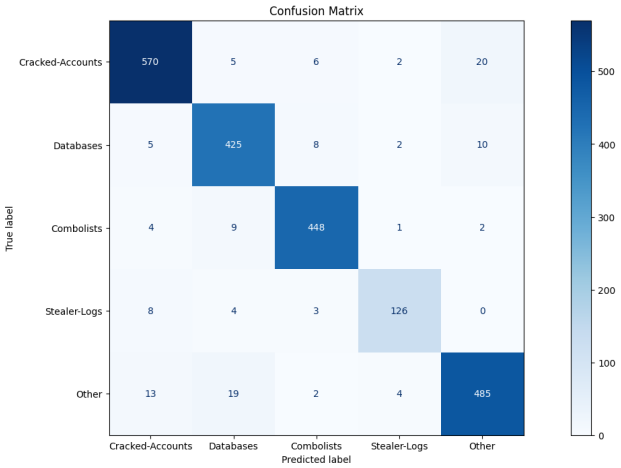


Figure 8. Confusion matrix

## 4.2. Retrieval-augmented generation

Figure 7 illustrates the evaluation process for the RAG results. We initially prepared an identical set of reference data and then used GPT 3.5 as the test LLM. We divided the approach into two configurations: the first involved GPT 3.5 using the RAG (our approach), and the second involved not using RAG. The key difference lies in our approach's ability to vectorize the reference data and then match it with the prompt to identify the most relevant fragment for generating a new prompt. The without RAG approach involves incorporating all reference data directly into the prompt. Theoretically, the without RAG approach should perform better since the LLM has access to all reference materials; however, this also results in the complete disclosure of our data to the LLM company. Our RAG approach aims to closely match the results of the without RAG approach by vectorizing our reference data to find the closest match to the prompt and subsequently forming a new prompt. Thus, the LLM company can not directly access our reference data, making this approach more suitable for organizations concerned with privacy.

Following the evaluations with and without RAG, two outputs were obtained: the response output and the number of prompt tokens. We then used GPT 4 to assess the scores of both responses. For this assessment, we used the without RAG approach with GPT 4, allowing the LLM to comprehend the complete dataset and score each response on a scale from 0 to 10 based primarily on the provided reference data. Ultimately, using eight prompts related to information leaks,

TABLE 5. RAG SCORE AND TOKEN RESULTS

| Approach | Mean Score (out of 10) | Mean Token Number |
|---|---|---|
| With RAG | 8.375 | 751 |
| Without RAG | 7.750 | 4723 |

we used two metrics: the scores and token numbers for both with and without RAG.

We organized eight questions related to information security and data breaches to evaluate responses with and without RAG. As well as text input, this system also accepts voice input, primarily implemented through the SpeechRecognition repository.

The average scores calculated across eight prompts yielded the results listed in Table 5.

In this experiment, our approach (with RAG) achieved a score of 8.375 out of 10, which is significantly better than 7.75 out of 10 achieved without using RAG.

The average number of tokens used for the eight prompts listed in Table 5

For an LLM, processing text primarily involves breaking it down into tokens. A higher number of tokens indicates a more complex process and, for LLM companies, directly impacts the cost, making a lower token count preferable.

Thus, the evaluation results suggest that our approach generates a greater response, also the with RAG approach significantly outperforms the without RAG approach in terms of token usage. Additionally, using RAG allows for the retention of organizational privacy by not uploading all reference data, thus safeguarding sensitive information.

## 5. Discussion

Our methodology integrates advanced monitoring and analysis techniques to address data leaks on the dark web efficiently:

1) Key finding: Through the implementation of Dark Watchdog in this paper, we can monitor data leaks involving companies on the dark web in real-time and report them immediately. This prevents situations where exposed data remains unnoticed for an extended period.
2) BERT Model Utilization: We use a BERT model to categorize posts related to leak events. The model is capable of identifying five distinct types of leak events, demonstrating versatility across various data types. This model can be applied to data from multiple forums, such as Dread [18] or X(formerly Twitter) [19], and others, enhancing our ability to track and analyze data across different platforms. The performance is better than other models.
3) Querying with Natural Language: We can interact with an LLM using natural language to comprehend the daily intelligence extracted from the dark web. This approach allows us to process and understand complex data patterns and security threats effectively.
4) Maintain Privacy with RAG: By using the RAG, we avoid exposing all our reference data to the LLM company. This approach not only preserves the privacy of our valuable reference data but also maintains the integrity and confidentiality of our monitoring processes.
5) Efficiency and Cost Reduction: The use of RAG significantly reduces the number of tokens required, which directly lowers our costs with the LLM company. Despite the reduction in token usage, with RAG approach performs comparable to without RAG that use more extensive data sets. This efficiency ensures that we can maintain high analytical standards while optimizing operational costs.

## 6. Conclusion

We utilized data from forums on the dark web to train a well-performing classification model that can be applied to various types of posts. The methods proposed in this paper have been tested and have yielded good results. In addition to performing well, using RAG also preserves user privacy by not requiring data to be disclosed to external parties.

These integrated methodologies enable us to swiftly and effectively identify potential data leaks, providing a robust defense mechanism against the exploitation of sensitive data on the dark web.

In the future, more expert opinions can be gathered to ask meaningful questions, such as whether the leak incidents are related to organizations, who is selling the data, and how many times the sold data has appeared. By creating prompts based on the questions a CTI report is concerned with, CTI reports can be automated in the future, and the dark web content that cybersecurity experts care about can accelerate the report creation process.

## References


[1] T. C. I. Team, "Double trouble: Ransomware with data leak extortion, part 1," https://www.crowdstrike.com/blog/double-trouble-ransomware-data-leak-extortion-part-1/.

[2] "How data breaches happen how to prevent data leaks," https://www.kaspersky.com/resource-center/definitions/data-breach.

[3] "Tor project: Anonymity online." [Online]. Available: https://www.torproject.org/

[4] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker, "Shining light in dark places: Understanding the tor network," in *International Symposium on Privacy Enhancing Technologies Symposium, vol.5134*, 2008, pp. 63–76.

[5] "Halloware Ransomware on Sale on the Dark Web for Only $40." [Online]. Available: https://www.bleepingcomputer.com/news/security/halloware-ransomware-on-sale-on-the-dark-web-for-only-40/

[6] "Scammer Uses Fake Tor Browser to Lure Victims to Supposed Dark Web Marketplace." [Online]. Available: https://www.bleepingcomputer.com/news/security/scammer-uses-fake-tor-browser-to-lure-victims-to-supposed-dark-web-marketplace/

[7] "Hackers' private chats leaked in stolen WeLeakData database." [Online]. Available: https://www.bleepingcomputer.com/news/security/hackers-private-chats-leaked-in-stolen-weleakdata-database/

[8] "Hacker Leaks 900 Enterprise VPN Server Passwords on Dark Web." [Online]. Available: https://healthitsecurity.com/news/hacker-leaks-900-enterprise-vpn-server-passwords-on-dark-web

[9] A. Zenebe, M. Shumba, A. Carillo, and S. Cuenca, "Cyber threat discovery from dark web," *EPiC Series in Computing*, vol. 64, pp. 174–183, 2019.

[10] H. Lawrence, A. Hughes, R. Tonic, and C. Zou, "D-miner: A framework for mining, searching, visualizing, and alerting on darknet events," in *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2017, pp. 1–9.

[11] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "Blackwidow: Monitoring the dark web for cyber security information," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900. IEEE, 2019, pp. 1–21.

[12] Y. H. Z. X. L. H. L. F. ZHENG Xianchun, WANG Rui, "A high performance tor web content monitoring system based on distributed crawlers," *Journal of Cyber Security*, vol. 8, no. 1, pp. 144–153, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "Darkbert: A language model for the dark side of the internet," *arXiv preprint arXiv:2305.08596*, 2023.

[16] M. Sultana, A. Taylor, L. Li, and S. Majumdar, "Towards evaluation and understanding of large language models for cyber operation automation," in *2023 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2023, pp. 1–6.

[17] J. Vijayan, "Leak site breachforums springs back to life weeks after fbi takedown," https://www.darkreading.com/cyberattacks-data-breaches/leak-site-breachforums-springs-back-to-life-weeks-after-fbi-takedown.

[18] "Dread." [Online]. Available: dreadytofatroptsdj6io7l3xptbet6onoyno2yv7jicoxknyazubrad.onion

[19] "X (formaly twitter)." [Online]. Available: https://x.com/

# Appendix A.
# RAG evaluation

TABLE 6. PROMPTS USED IN THE EXPERIMENT

| | |
|---|---|
| **Prompt 1** | As an information security analysis expert, please answer the following question: Please summarize the content. |
| **Prompt 2** | As an information security analysis expert, please answer the following question: What specific types of information are involved in this data? For example, are they account details, personally identifiable information, or something else? |
| **Prompt 3** | As an information security analysis expert, please answer the following question: What events are particularly important and require special attention? list 5 events |
| **Prompt 4** | As an information security analysis expert, please answer the following question: Please help me compile the occurrence of each country. |
| **Prompt 5** | As an information security analysis expert, please answer the following question: Are the data breach incidents concentrated in specific regions or countries? |
| **Prompt 6** | As an information security analysis expert, please answer the following question: Which specific industries or sectors are most affected by these leaks? |
| **Prompt 7** | As an information security analysis expert, please answer the following question: Please list the data that related to Asia |
| **Prompt 8** | As an information security analysis expert, please answer the following question: Please list the data that related to Finance |

TABLE 7. SAMPLE OF PROMPTS AND OUTPUT RESULTS USED FOR EVALUATING THE RAG (PROMPT 1)

| | |
|---|---|
| **Prompt** | As an information security analysis expert, please answer the following question: Please summarize the content |
| **Response (with RAG)** Score: 8/10 | The content provided seems to be related to leaked data from various sources including databases, cracked accounts, and other security breaches. The leaks include information such as user data from different countries and organizations, vulnerabilities in political and bank websites, and requests for sensitive data access. The data ranges from user credentials to personal information and the breaches involve institutions from different countries like the United States, Canada, Indonesia, Italy, and more. |
| **Response (without RAG)** Score: 7/10 | The content of the articles in the forum mainly consists of leaked data from various sources, including databases, combolists, and cracked accounts. There are also discussions about potential vulnerabilities, exploits, and information related to cyberattacks. One specific article mentions the leak of 12K Taiwan private combolists. |

TABLE 8. SAMPLE OF PROMPTS AND OUTPUT RESULTS USED FOR EVALUATING THE RAG (PROMPT 3)

| | |
|---|---|
| **Prompt** | As an information security analysis expert please answer the following question: What events are particularly important and require special attention? list 5 events |
| **Response (with RAG)** Score: 10/10 | Based on the provided context here are 5 important events that may require special attention from an information security analysis perspective: 1. United States Army Aviation and Missile Command Leaked - Download! - This event could have significant implications for national security and military operations. 2. China Huge Combolist Email: Password (2023/12 leak) 78 Millions Leaked - The large-scale leak of email and password combinations could lead to widespread privacy breaches and cyberattacks. 3. 2024 Qatar National Bank Database - The exposure of a bank database could pose risks to financial institutions and their customers. 4. Gov mails [Access any law enforcement portal and request Data] - Breaches involving government emails and law enforcement portals may have serious consequences for national security and public safety. 5. WTB PayPal API keys STRIPE API keys IN BULK - The potential sale of API keys in bulk could result in unauthorized access to sensitive payment information and financial transactions. |
| **Response (without RAG)** Score: 7/10 | Events that are particularly important and require special attention in the context of information security analysis are: 1. Cyberattack on Philippines National Police 2. China Huge Database Leaks (including Chinese citizens companies and financial investors) 3. 200K UrlUserPass Base Leak 4. SQL DB - UK FREE leak 5. United States Army Aviation and Missile Command Leaked |